Fiabilité des Mémoires Non-Volatiles de type Flash en architectures NOR et NAND

Jérémy Postel-Pellerin

ANNEE : 2008

THESE

présentée devant

L'UNIVERSITE DE PROVENCE (AIX-MARSEILLE I)

par

Jérémy POSTEL-PELLERIN
Ingénieur Ecole Centrale Marseille cursus ENSPM

pour obtenir

LE GRADE DE DOCTEUR
Ecole Doctorale 353 : Physique, Modélisation et Sciences pour l'Ingénieur

---

Fiabilité des Mémoires Non-Volatiles de type Flash en architectures NOR et NAND

Directeur de thèse : Frédéric LALANDE
Co-Directeur de thèse : Pierre CANET

Soutenue le 08/12/2008 devant la commission d'examen :

Rapporteurs :

Mireille COMMANDRE	**Professeur**	**Ecole Centrale Marseille**
Yves DANTO	**Professeur Emérite**	**Université de Bordeaux I**

Examinateurs :

Pascal MASSON	**Professeur**	**Université de Nice**
Frédéric LALANDE	**Professeur**	**Université Paul Cézanne**
Pierre CANET	**Professeur**	**Université de Provence**
François JEULAND	**Manager**	**ATMEL-Rousset**

Remerciements

Je tiens à remercier en premier lieu le laboratoire IM2NP et la société ATMEL qui m'ont permis, en me confiant conjointement ce travail, d'effectuer ma thèse dans les meilleures conditions techniques et scientifiques.

Je remercie tout particulièrement Monsieur Rachid Bouchakour, Directeur de l'IM2NP, ainsi que Monsieur François Jeuland, Manager de l'équipe Device Engineering, de m'avoir accueilli au sein de leurs effectifs.

Au sein du laboratoire, je tiens à remercier très chaleureusement Messieurs Pierre Canet, Professeur et Frédéric Lalande, Professeur, pour leurs précieux conseils tant scientifiques qu'humains depuis déjà toutes ces années.

Je voudrais également remercier Madame Laurence Morancho, Messieurs Bernard Bertello et Gilles Festes d'avoir suivi mon travail au jour le jour, d'avoir su m'aider à cadrer mon étude et de m'avoir apporté leurs connaissances techniques précieuses pour ma thèse.

Madame Mireille Commandré m'a fait l'honneur d'accepter d'évaluer mon travail en qualité de rapporteur et je l'en remercie très vivement.

Que Monsieur Yves Danto soit également remercié d'avoir bien voulu prendre de son temps pour juger de ma thèse en tant que rapporteur.

J'aimerais dire un grand merci à l'ensemble des équipes de travail de l'IM2NP et d'ATMEL que j'ai eu la joie de côtoyer lors de cette thèse et qui ont très bien su m'intégrer par leur gentillesse, leur bonne humeur et leur aide.

Merci donc à toute l'équipe Device Engineering, François, Laurence, Bernard, Gilles F., Bruno, Jean-Yves, Stéphane, Gilles L., Patrick, Cathy, Elsa, Gilles M., Thibault, Jean-Marc B., Yves et Jean-Marc D.

Un grand merci également à tous les thésards ou anciens thésards de l'IM2NP : Valéry, Juliano et Arnaud qui m'ont appris le métier, Marraine Manon, Julien, Ramos et Daniele de la salle ATMEL, Patrick, Jean-René, Anthony, Fabrice, Thibaut, Marc, Yannick, Emmanuel, Jérémy, Samir, Amir, Olivier, Fayrouz, Oussama, Rémy et tous ceux que j'ai maladroitement oubliés.

Enfin, merci à ma petite femme Marie pour s'être occupé de moi pendant cette longue période de travail et à notre bébé Flavie de m'avoir laissé dormir un peu entre les journées de travail ! Merci également aux parents de Marie d'être venus nous soulager dans nos nombreuses tâches quotidiennes. Merci à mes parents qui se sont parfois sacrifiés pour me permettre de poursuivre mes études et de suivre ma voie.

A Marie
A Flavie

Table des matières

Liste des Symboles et des Abréviations

Symboles	Signification	Unité
BL	Bit Line (ligne de bit)	-
CG	Control Gate (grille de contrôle)	-
CHE	Channel Hot Electrons (injection d'électrons chauds)	-
CMOS	Complementary Metal Oxide Semiconductor	-
CNT	Carbon NanoTube	-
CRAM	Chalcogenide Random Access Memory	-
DRAM	Dynamic Random Access Memory	-
EEPROM	Electrically Erasable and Programmable Read Only Memory	-
EPROM	Electrically Programmable Read Only Memory	-
FeRAM	Ferroelectric Random Access Memory	-
FG	Floating Gate (grille flottante)	-
FLOTOX	FLOating gate Thin OXide	-
FN	Fowler-Nordheim	-
GIDL	Gate Induced Drain Leakage current	-
MNOS	Metal Nitrure Oxide Semiconductor	-
MNV	Mémoire Non Volatile	-
MOS	Metal Oxide Semiconductor	-
MTJ	Magnetic Tunnel Jonction	-
NVM	Non Volatile Memory	-
ONO	Oxyde Nitrure Oxyde	-
OUM	Ovonic Unified Memory	-
PCM	Phase Change Memory	-
PolySi	PolySilicium	-
PRAM	Phase-change Random Access Memory	-
PROM	Programmable Read Only Memory	-
PZT	Titano-Zirconiate de Plomb	-
PZTN	Titano-Zirconiate de Plomb et Niobium	-
Sentaurus P	Sentaurus Process (simulateur procédé)	-
Sentaurus SE	Sentaurus Structure Editor (éditeur de structure)	-
Sentaurus D	Sentaurus Device (simulateur électrique)	-

SILC	Stress Induced Leakage Current	-
SIMOS	Stacked Injection MOS	-
SiO2	Dioxyde de Silicium	-
SONOS	Silicium Oxyde Nitrure Oxyde Silicium	-
SRAM	Static Random Access Memory	-
Stress	Contrainte électrique	-
TMR	Tunnel MagnetoResistance	-
TPFG	Textured Poly Floating Gate	-
WL	Word Line (ligne de mot)	-
C_{CGy}	Capacité de couplage entre les grilles de contrôle de deux cellules voisines, situées sur la même ligne de bit	F
C_{FE}	Capacité FerroElectrique de stockage de l'information dans la cellule mémoire FeRAM	F
C_{FGx}	Capacité de couplage entre les grilles flottantes de deux cellules voisines, situées sur la même ligne de mot	F
C_{FGxy}	Capacité de couplage entre les grilles flottantes de deux cellules voisines en diagonale	F
C_{FGy}	Capacité de couplage entre les grilles flottantes de deux cellules voisines, situées sur la même ligne de bit	F
C_{FG-CG}	Capacité de couplage entre la grille flottante d'une cellule et la grille de contrôle d'une cellule voisine, située sur la même ligne de bit	F
C_{pp}	Capacité inter-poly entre la grille flottante et la grille de contrôle	F
ϕ_0	Hauteur de barrière Si/SiO2	eV
ϕ_B	Potentiel de substrat	eV
I_{FN}	Courant tunnel Fowler-Nordheim	A
Q_{FG}	Quantité de charges stockées dans la grille flottante	C
S_{pp}	Surface de la capacité inter-poly	m^2
t_{pp}	Epaisseur de la capacité inter-poly	m
V_{cg}	Potentiel de grille de contrôle	V
V_{fg}	Potentiel de grille flottante	V
V_{ms}	Potentiel métal/semiconducteur	V
V_T	Tension de seuil	V
V_{T0}	Tension de seuil naturelle (en l'absence de charges dans la grille flottante)	V
$V_{T_{Bas}}$	Tension de seuil basse due à la présence de charges positives dans la grille flottante	V
$V_{T_{Haut}}$	Tension de seuil élevée due à la présence de charges négatives dans la grille flottante	V
ϵ_0	Permittivité diélectrique du vide	$F.m^{-1}$
ϵ_{Si}	Permittivité diélectrique relative du silicium	-
ϵ_{SiO2}	Permittivité diélectrique relative du dioxyde de silicium	-
q	Charge élémentaire de l'électron	C

Introduction Générale

Pour satisfaire aux impératifs économiques et améliorer les performances de ses produits, l'industrie de la microélectronique est amenée à réduire en permanence les dimensions de ses dispositifs élémentaires (transistors ou cellules mémoires). A chaque génération, les acteurs de la microélectronique sont confrontés à de nouveaux défis techniques et scientifiques qui doivent être résolus au plus vite par souci de compétitivité. La fiabilité des Mémoires Non-Volatiles (ou MNV) telles que les mémoires *Flash* est un de ces points primordiaux pouvant mettre en jeu la sécurité du client, en particulier dans le domaine de l'automobile. Ce travail de thèse a été réalisé en collaboration entre la société ATMEL Corporation sur le site de Rousset et l'Institut Matériaux Microélectronique Nanosciences de Provence (*IM2NP*, UMR CNRS 6242), dans la cadre d'une convention CIFRE et porte sur une étude de la fiabilité des Mémoires Non-Volatiles de type Flash en architectures NOR et NAND.

Ce manuscrit, divisé en quatre chapitres, propose une étude de différents aspects de la fiabilité des Mémoires Non-Volatiles et plus particulièrement des Mémoires Flash.

Le premier chapitre présente une étude bibliographique des différentes mémoires à semiconducteurs existantes, en particulier non-volatiles, afin de situer le cadre de cette thèse qui porte sur les MNV de type Flash. Après une brève description des mémoires volatiles Static Random Access Memory et Dynamic Random Acces Memory, nous donnerons une vue d'ensemble du fonctionnement et du marché des mémoires non-volatiles actuelles. Le domaine de la microélectronique étant en continuelle évolution, nous présenterons également les mémoires émergentes en cours de développement, censées remplacer à court ou moyen terme les mémoires actuelles. Nous détaillerons ensuite les différentes architectures de matrices mémoires Flash qui seront étudiées dans nos travaux, ainsi que les types de perturbations qu'elles peuvent subir lors de leur fonctionnement.

Dans le second chapitre, nous présentons une étude des méthodes de programmation dans l'objectif de diminuer les dégradations lors des successions de phases de programmation et d'effacement des cellules mémoires Flash. Plusieurs voies d'amélioration sont explorées en jouant sur les signaux utilisés. Nous nous basons tout d'abord sur la littérature montrant la possibilité d'améliorer nettement la tenue des

cellules au test en endurance (cyclage) en utilisant des signaux de très courte durée, de l'ordre de quelques centaines de nanosecondes. Nous appliquons cette méthode sur deux de nos structures Flash en architecture NAND. Une autre voie de réduction des dégradations des cellules, subies lors des phases de programmation et d'effacement, repose sur l'optimisation de la forme des signaux utilisés dont nous proposons dans la suite du chapitre une étude théorique. Nous développons ensuite un algorithme de programmation intelligente qui nous permet de garantir les tensions de seuil en programmation et en effacement quasi-constantes en fonction du nombres de cycles. Cela nous affranchit des décalages de V_T causés par le piégeage de charges dans l'oxyde tunnel, au détriment de la durée d'effacement qui augmente avec le nombre de cycles.

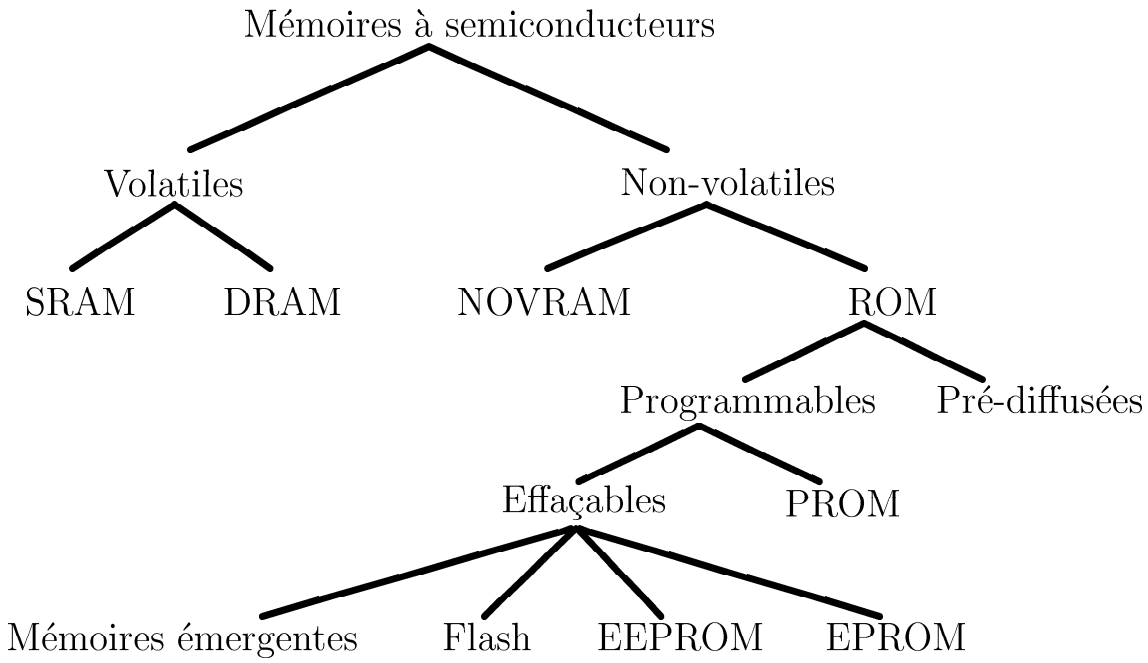
Afin de réduire les coûts de production des mémoires Flash, l'idée de stocker plusieurs bits d'information sur une seule cellule mémoire est apparue, c'est ce que l'on appelle le stockage multi-niveaux. Cependant, de nouveaux problèmes de discrimination des états se posent, du fait de la présence de quatre niveaux au lieu des deux niveaux habituels "0" et "1". **Le troisième chapitre** étudie les mécanismes de fuites de cellules mémoires Flash en architecture NOR multi-niveaux lors de mesures en rétention à différentes températures, destinées à fournir une estimation de la durée de stockage de l'information à température ambiante. Nous présentons dans un premier temps le protocole de l'expérience, notamment les résultats de cyclage avant rétention, puis nous exposerons les résultats en rétention avant de proposer une modélisation des pertes de charges par superposition de deux mécanismes. Nous apportons aussi une modélisation à un phénomène de "gain de charges", conduisant à une augmentation de la tension de seuil lors du test en rétention.

Le dernier chapitre traite des perturbations pouvant intervenir à l'intérieur des matrices mémoires, causées soit par les polarisations appliquées, soit par les capacités de couplage parasite entre cellules. Nous étudions les perturbations de grille sur cellule mémoire Flash en architectures NOR puis NAND avant et après cyclage avant d'aborder le problème de dégradation des cellules inhibées lors du cyclage de la cellule sélectionnée. Pour expliquer cette dégradation, nous combinons des mesures sur silicium à des simulations TCAD bidimensionnelles, complétées par une prise en compte des capacités parasites de couplage tri-dimensionnelles dans la matrice mémoire, dont l'extraction est faite par simulation TCAD, par mesure et par calcul géométrique. Nous proposons enfin une étude de l'impact de paramètres process tels que les implants de source et drain ou des dimensions entre cellule et transistor de sélection, sur le phénomène de channel boosting.

Nous concluons enfin de façon générale sur le travail effectué lors de ces trois années et exposons les différentes perspectives qui pourraient être pertinentes pour la suite de ce sujet.

Chapitre 1

Etat de l'art de la technologie des mémoires



Sommaire

Ce premier chapitre a pour but de donner une vision d'ensemble des différents types de mémoires à semiconducteurs.

Avant d'entrer dans les détails des mémoires non-volatiles, nous présentons tout d'abord une classification des mémoires à semiconducteurs existantes. Les mémoires volatiles (SRAM[1], DRAM[2]) sont traitées rapidement tandis que les mémoires non-volatiles (EEPROM[3], Flash, ...) sont davantage détaillées, l'ensemble de ce manuscrit portant sur des mémoires non-volatiles. La tendance actuelle est la recherche d'une mémoire "universelle" qui serait à la fois non-volatile, ayant un faible coût, avec des vitesses de programmation et d'effacement élevées, dont l'endurance et la rétention seraient les plus grandes possibles, permettant une forte intégration, tout cela en devant rester compatible avec les technologies existantes (CMOS[4], ...). Nous comparerons ainsi tout particulièrement les caractéristiques de fonctionnement de ces différentes mémoires non-volatiles, chacune possédant ses propres avantages et inconvénients dans la recherche de cette mémoire idéale.

Nous décrivons par la suite les modes de fonctionnement d'une cellule Flash, aussi bien en architecture NOR qu'en architecture NAND, architectures qui sont décrites puis comparées en vue de leur étude dans les chapitres suivants. Nous terminons alors par une étude bibliographique des perturbations pouvant intervenir lors de l'utilisation de la cellule mémoire dans chacune de ces architectures.

[1]Static Random Access Memory
[2]Dynamic Random Access Memory
[3]Electrically Erasable Programmable Read Only Memory
[4]Complementary Metal Oxide Semiconductor

1.1 Les différents types de mémoires à semiconducteurs

Les mémoires à semiconducteurs peuvent être classées en mémoires volatiles ou non-volatiles, programmables ou non et effaçables ou non, ce qu'illustre la figure 1.1.

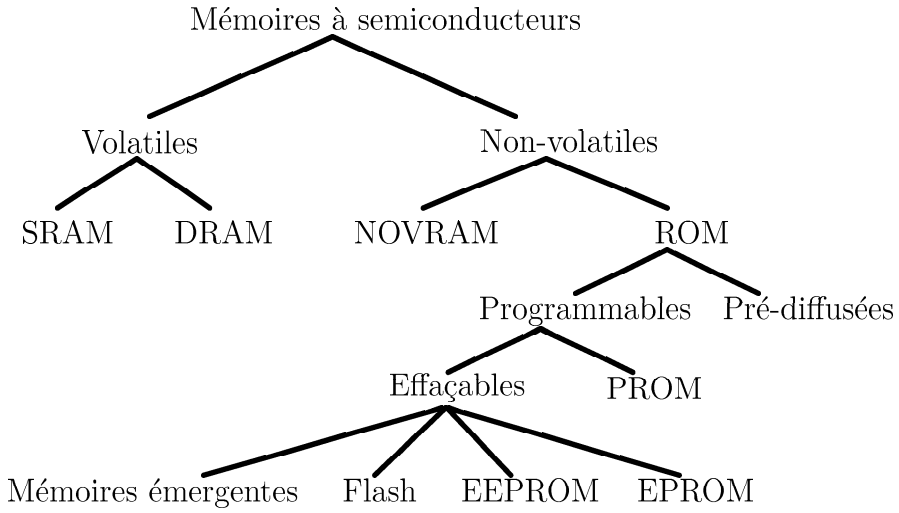


FIG. 1.1 – Classification des différents types de mémoires à semiconducteurs.

1.1.1 Les mémoires volatiles

Les mémoires volatiles doivent leur nom au fait qu'en l'absence d'alimentation électrique elles perdent l'information. Les deux principaux types de mémoires volatiles sont la SRAM[5] et la DRAM[6].

1.1.1.1 SRAM

Le terme **"Static"** RAM indique que la mémoire retient l'information tant que l'alimentation est maintenue, ce qui n'est pas le cas de la **"Dynamic"** RAM comme nous le verrons dans le paragraphe suivant. La mémoire SRAM comprend en général six transistors (Figure 1.2), c'est la mémoire à semiconducteurs la plus rapide mais en contrepartie, son coût est relativement élevé du fait de sa plus faible densité, ce qui limite ses applications dont la principale est la mémoire cache des ordinateurs pour laquelle elle est parfaitement adaptée.

[5]Static Random Access Memory
[6]Dynamic Random Access Memory

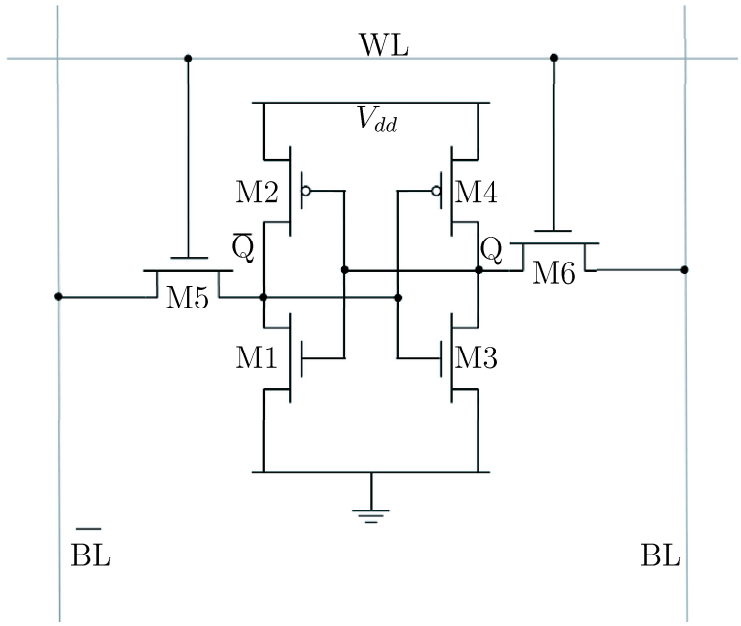


FIG. 1.2 – Cellule mémoire SRAM à six transistors.

1.1.1.2 DRAM

Pour étendre les domaines d'application des mémoires SRAM, il faut une mémoire ayant un plus faible coût, ce qui est le cas de la DRAM qui peut être réalisée à l'aide d'un seul transistor et d'une capacité, comme le montre la figure 1.3. On parle alors de 1T1C-DRAM. La densité d'intégration est alors beaucoup plus importante que pour la SRAM mais du fait des fuites de la capacité, l'information doit être rafraîchie régulièrement. Egalement, lors de la lecture de la donnée stockée, l'information est perdue par décharge de la capacité dans la ligne de bit et doit donc au préalable être recopiée dans un circuit annexe puis réécrite après la lecture.

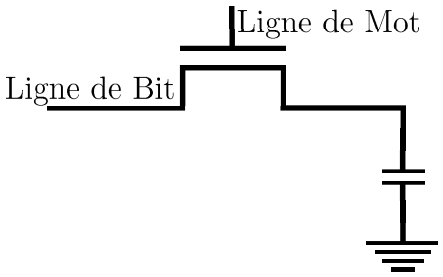


FIG. 1.3 – Cellule mémoire DRAM à un transistor et une capacité, ou 1T1C-DRAM.

1.1.2 Technologies des mémoires non-volatiles

Contrairement aux mémoires volatiles, les mémoires non-volatiles présentent la particularité de pouvoir garder l'information stockée même en l'absence d'alimentation électrique. On imagine aisément l'attrait de ces mémoires pour de nombreuses applications, telles que la sauvegarde de données dans les téléphones portables, la photographie numérique, les clés USB, ...

Les mémoires non-volatiles reposent sur plusieurs grands types de technologies.

1.1.2.1 Technologies à grille flottante

La solution la plus répandue pour lutter contre la volatilité est l'utilisation de transistors MOS, dont on décale la tension de seuil par une charge stockée dans une grille isolée au-dessus du canal. Les technologies à grille flottante consistent à ajouter, entre la grille et le canal d'un transistor MOS traditionnel, une seconde grille en matériau conducteur ou semiconducteur afin d'y isoler des charges, dans le but de faire varier la tension de seuil du transistor. La plupart du temps, les charges sont injectées à travers un diélectrique, en général un oxyde, le plus courant étant le dioxyde de Silicium SiO2, placé entre la grille flottante et le canal du transistor, comme le représente la figure 1.4. Enfin, les deux grilles de ce "transistor" sont séparées par un oxyde, le plus commun étant un oxyde nitruré Oxyde-Nitrure-Oxyde (ONO).

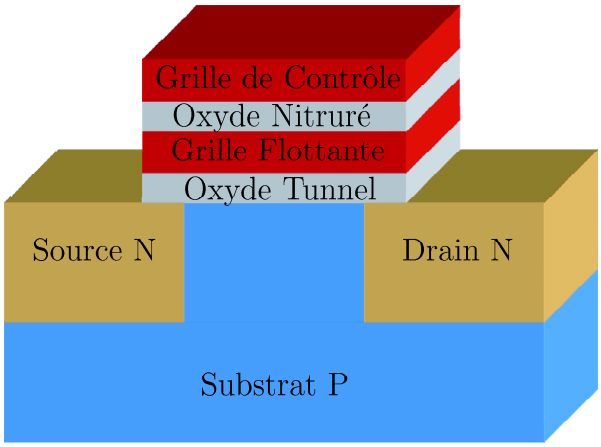


FIG. 1.4 – Schéma d'une structure à grille flottante.

On distingue les différentes technologies à grille flottante selon le mécanisme utilisé en programmation.

La technologie SIMOS

Dans le cas de la technologie SIMOS[7], le mécanisme utilisé lors de la programmation de la cellule mémoire est le phénomène d'injection d'électrons chauds dans la

[7]Stacked Injection MOS

grille flottante à travers l'oxyde de grille situé en-dessous. La figure 1.5 schématise ce mécanisme dans une cellule dont les deux grilles sont en Polysilicium, l'oxyde de grille est en SiO2 et l'oxyde entre les deux grilles est un oxyde nitruré ONO.

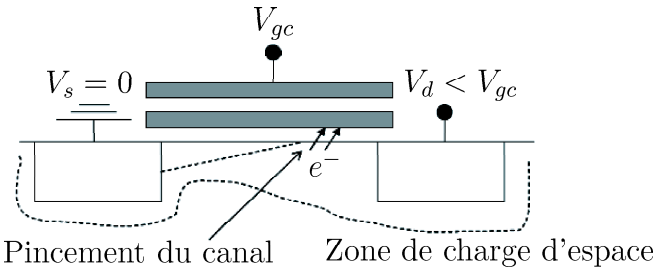


FIG. 1.5 – Injection d'électrons chauds en programmation dans la technologie SIMOS.

Dans cette technologie, l'efficacité de l'injection d'électrons chauds, et par conséquent l'efficacité de la programmation, dépendent fortement du dopage de substrat, de la longueur effective du canal et de la longueur de la zone de recouvrement entre la grille flottante et le drain [Brown'98].

La technologie FLOTOX

Dans le cas de la technologie FLOTOX[8], le mécanisme utilisé lors de la programmation de la cellule mémoire est le courant tunnel Fowler-Nordheim qui permet l'injection d'électrons dans la grille flottante à travers l'oxyde de grille. En utilisant cette technologie FLOTOX, il est possible de créer des structures à un ou à deux niveaux de silicium polycristallin, appelées respectivement structures "Simple-Poly" (Figure 1.6 (a)) et "Double-Poly" (Figure 1.6 (b)).

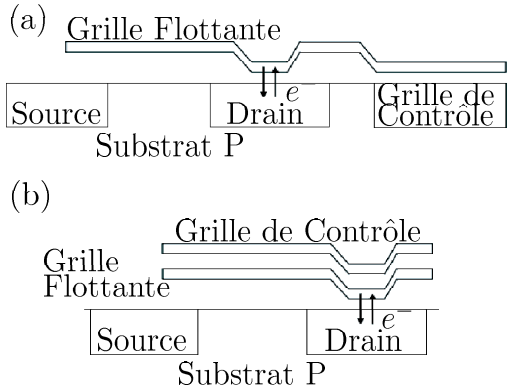


FIG. 1.6 – Structures en technologie FLOTOX : (a) Simple-Poly et (b) Double-Poly.

[8]FLOating gate Thin OXide

Dans les structures Simple-Poly, le seul niveau de silicium polycristallin forme la grille flottante, tandis que la grille de contrôle est enterrée dans le substrat et séparée de la grille flottante par une couche d'oxyde d'une épaisseur de l'ordre de 20 à 30 nm.

En ce qui concerne les structures Double-Poly, un niveau constitue la grille flottante tandis que le second niveau forme la grille de contrôle. Ces deux grilles sont séparées par une couche d'oxyde, le plus souvent un ONO d'une vingtaine de nanomètres, appelé "oxyde interpoly".

Dans les deux types de structures, la grille flottante est séparée du drain par un oxyde mince, de 7 à 10 nm environ, permettant l'injection Fowler-Nordheim.

La technologie TPFG

Comme pour la technologie FLOTOX, la technologie TPFG[9] utilise l'effet tunnel Fowler-Nordheim en programmation, mais la couche à travers laquelle les électrons transitent pour aller dans la grille flottante est en poly-oxyde. Au lieu de faire croître la couche d'oxyde sur un silicium monocristallin, on fait croître thermiquement cet oxyde sur une couche de silicium polycristallin. La présence d'aspérités à l'interface polysilicium/polyoxyde permet d'utiliser des champs électriques moins intenses pour injecter les électrons dans la grille flottante. Cette réduction des champs électriques appliqués induit une dégradation moins importante et par conséquent une amélioration de la fiabilité.

Le tableau 1.1 résume l'ensemble des technologies à grille flottante ainsi que leurs mécanismes de programmation et les différentes mémoires basées sur ces technologies. Les mémoires EPROM[10], EEPROM[11] et Flash-EEPROM seront présentées ultérieurement dans ce manuscrit en partie 1.1.3.

Technologie	**Mécanisme de programmation**	**Mémoires utilisant cette technologie**
SIMOS	Injection par électrons chauds	EPROM, Flash
FLOTOX	Injection par effet tunnel Fowler-Nordheim à travers un oxyde mince (7 à 10 nm)	EEPROM, Flash
TPFG	Injection par effet tunnel Fowler-Nordheim à travers une couche en Poly-oxyde	EEPROM, Flash

Tab. 1.1 – Résumé des Technologies à grille flottante.

[9]Textured Poly Floating Gate
[10]Erasable Programmable Read Only Memory
[11]Electrically Erasable Programmable Read Only Memory

Nous pouvons remarquer que la technologie EPROM utilise nécessairement une injection d'électrons chauds lors de la programmation ce qui la classe dans la technologie SIMOS. En ce qui concerne les mémoires EEPROM, la programmation se fait dans tous les cas par effet tunnel Fowler-Nordheim, soit à travers un oxyde mince pour une technologie FLOTOX, soit à traverse une couche de poly-oxyde pour la technologie TPFG. La mémoire Flash-EEPROM peut utiliser chacun des trois mécanismes de programmation selon l'architecture considérée et donc se classer dans les trois catégories de technologies à grille flottante.

1.1.2.2 Technologies à piégeage de charges

Dans les technologies dites à piégeages de charges, la grille flottante en polysilicium est remplacée par une couche de nitrure Si_3N_4 qui joue le rôle de pièges pour les charges.

La technologie MNOS

Dans le cas de la technologie MNOS[12], les charges passent par effet tunnel du substrat via un oxyde fin de l'ordre de $1,5$ à 3 nm vers une couche de nitrure où elles sont piégées. Au dessus, une grille métallique est directement déposée sur le nitrure. La grille directement placée sur le nitrure peut poser des problèmes de passage de charges entre la grille et le nitrure.

La technologie SONOS

Pour remédier au problème d'injection de charges de la grille vers le nitrure, la technologie SONOS[13], illustrée par la figure 1.7 comporte un oxyde fin, de l'ordre de 2 à $3nm$ entre la grille, en polysilicium, et le nitrure [Chen'77].

Avec les technologies à piégeage de charges, on comprend "aisément" la possibilité de stocker plusieurs bits en utilisant plusieurs localisations de charges piégées, proche du drain et de la source par exemple. Ainsi, au lieu de stocker uniquement les niveaux "0" et "1", on peut ajouter deux états supplémentaires en distinguant si les charges sont stockées du côté de la source ou du côté du drain. Dans une seule cellule, on a alors quatre états possibles "00", "01", "10" et "11". Cela permet un gain énorme en coût car sur la même surface de silicium, on peut stocker deux fois plus d'informations. Cette technique consistant à mettre plusieurs états dans une seule cellule s'appelle du multi-niveaux[14] dont une étude sera proposée dans le chapitre 3 sur une mémoire Flash.

[12]Metal Nitrure Oxide Semiconductor
[13]Silicium Oxyde Nitrure Oxyde Silicium
[14]ou MLC, Multi-Level Cell

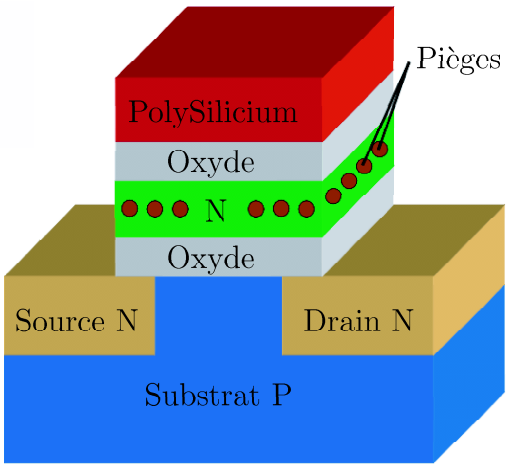


FIG. 1.7 – Cellule mémoire SONOS.

1.1.3 Les mémoires non-volatiles

A partir de ces différentes technologies de stockage de l'information, plusieurs mémoires non-volatiles ont été développées au cours des années.

1.1.3.1 ROM

Comme son nom l'indique, la mémoire ROM[15] est uniquement destinée à être lue. Sa programmation est effectuée une fois pour toutes lors de la fabrication par activation ou non de transistors par masquage. Les principales utilisations des mémoires ROM sont le stockage des jeux d'instructions des microprocesseurs ainsi que la sauvegarde des éléments vitaux à l'initialisation d'un ordinateur lors du démarrage.

1.1.3.2 PROM

Le principe de la mémoire PROM[16] est le même que pour une mémoire ROM à l'exception près que la programmation de la cellule s'effectue par l'utilisateur avec des structures à bases de fusibles. Ces fusibles sont grillés ou non par l'utilisateur en fonction de ses besoins. En revanche, comme pour la ROM, il n'est ensuite pas possible d'effacer ni de re-programmer cette cellule.

1.1.3.3 EPROM

Une évolution de la mémoire PROM a consisté à pouvoir l'effacer et l'écrire à volonté, on appelle cette mémoire EPROM[17]. Il est en effet possible de programmer

[15] Read Only Memory
[16] Programmable Read Only Memory
[17] Erasable Programmable Read Only Memory

une EPROM électriquement en quelques minutes pour une mémoire de quelques ko[18] par l'intermédiaire d'un programmateur spécial et par conséquent en ôtant la puce mémoire du circuit. Pour effacer une mémoire EPROM, il faut également ôter la puce du circuit mais l'effacement se fait par une exposition à un rayonnement Ultra-Violet via une fenêtre transparente située sur le boîtier de la puce.

1.1.3.4 EEPROM

La mémoire EPROM étant effaçable uniquement par exposition à un rayonnement Ultra-Violet pendant plusieurs minutes, une nouvelle mémoire a été développée, programmable et effaçable électriquement, la mémoire EEPROM[19]. La cellule mémoire de type EEPROM, dont la figure 1.8 présente une vue en coupe SEM[20], est une mémoire à double poly reposant le plus souvent sur la technologie FLOTOX. Elle est constituée d'un transistor à grille flottante appelé "transistor d'état" placé en série avec un transistor MOS appelé "transistor de sélection". Du fait de la nécessité de deux transistors pour stocker l'information, la mémoire EEPROM utilise une grande surface de silicium ce qui limite ses champs d'application.

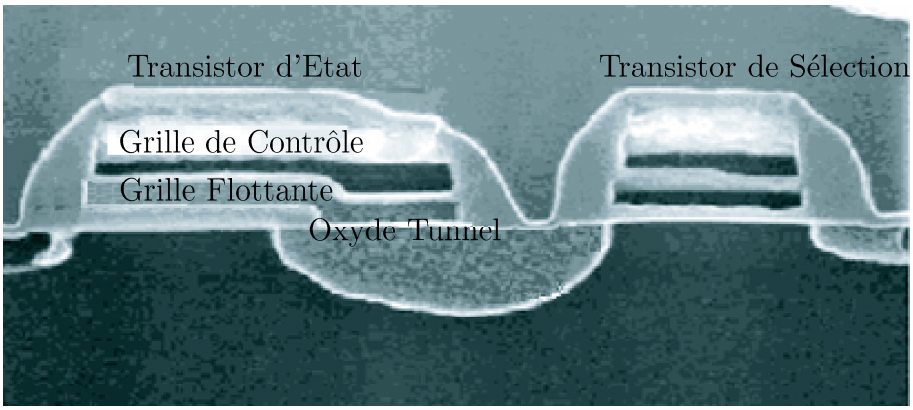


FIG. 1.8 – Cellule mémoire EEPROM.

1.1.3.5 Flash-EEPROM

La cellule mémoire Flash-EEPROM, appelée Flash dans la suite de ce manuscrit, a été développée à partir de la cellule mémoire EEPROM avec une volonté d'en améliorer la densité d'intégration. Ainsi, au lieu d'utiliser deux transistors pour stocker l'information, la cellule Flash, également effaçable électriquement mais par bloc, n'utilise que le transistor à grille flottante, appelé transistor d'état. Ce transistor peut être effacé par injection tunnel Fowler-Nordheim et programmé soit par

[18] 1 ko = 1 kilo octets = 1024 octets
[19] Electrically Erasable Programmable Read Only Memory
[20] Scanning Electron Microscopy

injection d'électrons chauds (ou **C**hannel **H**ot **E**lectron), soit par injection tunnel Fowler-Nordheim, comme décrit par la figure 1.9.

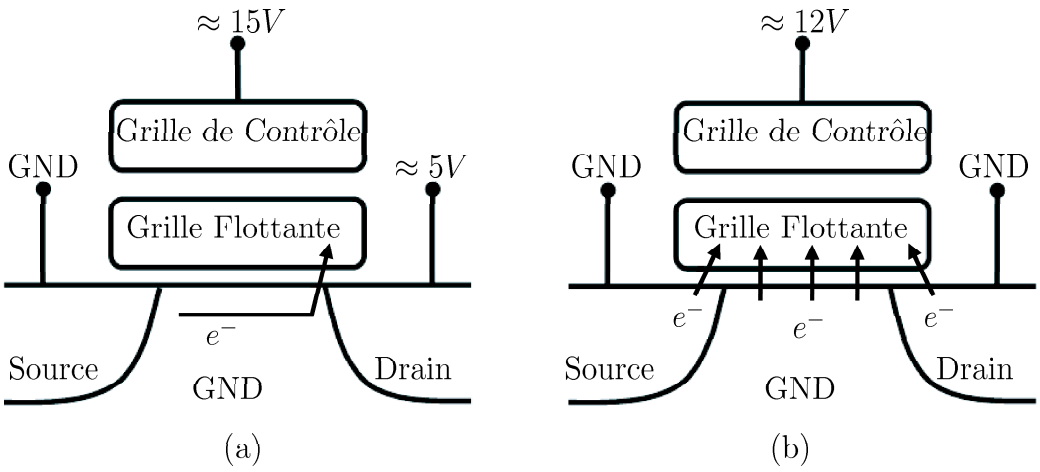


FIG. 1.9 – (a) Programmation d'une cellule Flash par injection d'électrons chauds (b) Programmation par injection tunnel Fowler-Nordheim.

1.1.4 Les mémoires émergentes

De nombreux axes de recherche sont actuellement exploités en vue de développer un nouveau type de mémoire qui allierait tous les avantages des mémoires présentées précédemment, à savoir notamment la non-volatilité, la rapidité de programmation et d'effacement et avec une forte endurance [Pavan'97]. Plusieurs types semblent particulièrement prometteurs et pourraient devenir la mémoire du futur : la MRAM[21], la FeRAM[22], les mémoires à changement de phase, les mémoires à nano-cristaux et la NRAM[23].

1.1.4.1 MRAM

Le développement de la MRAM a débuté dans les années 1990. Contrairement aux mémoires traditionnelles, l'information n'est pas stockée sous forme de charges électriques, mais sous forme magnétique. Chacune des cellules mémoires comporte deux éléments ferromagnétiques ayant chacun leur propre orientation magnétique. L'élément inférieur est dit "fixe" car son état est permanent et sa polarité spécifique. En revanche, l'élément supérieur est dit "libre" car il peut changer d'état par application d'un champ magnétique ou électrique extérieur. Ces deux éléments sont

[21]Magnetic Random Access Memory
[22]Ferroelectric Random Access memory
[23]Nano Random Access Memory, utilisant les nanotubes de carbone

séparés par une couche d'isolant tunnel. Le stockage des données repose sur la mesure de la résistance électrique de la cellule. En effet, d'après le phénomène de magnétorésistance[24] qui intervient sur une jonction tunnel magnétique[25] formée par un isolant entre deux couches ferromagnétiques, la résistance dans le tunnel varie avec l'orientation du champ magnétique entre les deux plaques. D'après la figure 1.10, si les deux éléments ferromagnétiques ont la même orientation magnétique, la résistance est faible et on stocke un "0" tandis que si les deux éléments ferromagnétiques ont des orientations opposées, la résistance est élevée et on stocke un "1".

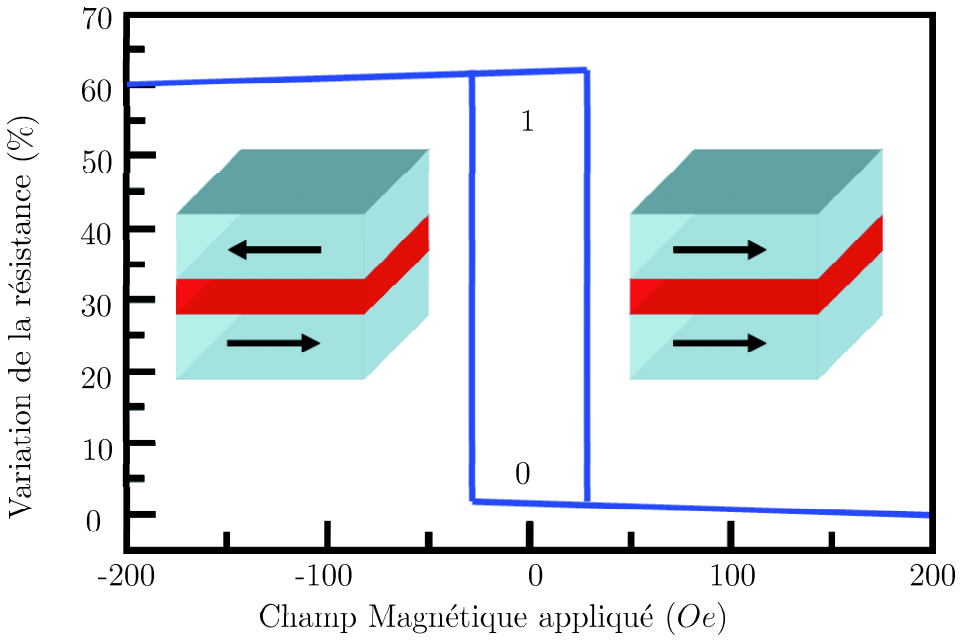


FIG. 1.10 – Cellule mémoire MRAM.

Pour écrire la cellule, il faut modifier l'orientation magnétique de la couche supérieure. La première méthode possible consiste à faire passer un fort courant ce qui génère le champ magnétique nécessaire à la modification de l'orientation. Cependant, ce fort courant pose un problème évident en terme de consommation, c'est pourquoi une seconde méthode d'écriture a été développée, le "Spin Transfer Switching". Cette technique repose sur l'envoi d'un nombre suffisant d'électrons de spins cohérents qui vont provoquer la repolarisation du matériau. La consommation est ainsi fortement diminuée. Subsistent néanmoins de nombreux problèmes liés à la présence d'un système qui maintient la cohérence entre les spins des électrons, d'où une perte de place, même si la densité reste meilleure que pour une mémoire SRAM.

[24] ou TMR, Tunnel MagnetoResistance
[25] ou MTJ, Magnetic Tunnel Jonction

Le tableau 1.2 résume les principales caractéristiques des mémoires MRAM. Les tailles de cellules sont exprimées en F^2, unité de surface correspondant à l'aire d'un carré de côté la "**F**eature size", soit la taille du nœud technologique. La taille du nœud technologique correspond à la plus petite dimension lithographiée sur la structure étudiée, en général pour les mémoires la longueur du canal de la cellule mémoire. La scalabilité traduit quant à elle la possibilité de réduire la taille de la technologie.

Mémoire MRAM	
Taille de cellule (F^2)	20
Endurance en écriture/lecture	$10^{16}/\infty$
Temps de lecture (aléatoire)	$30ns$
Temps de programmation (/octet)	$30ns$
Temps d'effacement (/octet)	$30ns$
Scalabilité	Faible
Limite de scalabilité	Densité de courant
Possibilité de multi-niveaux	Non
Coût relatif par bit	Elevé
Maturité	Faible

TAB. 1.2 – Résumé des caractéristiques des mémoires MRAM [ITRS'05a] [Pirovano'05].

1.1.4.2 FeRAM

La mémoire FeRAM ressemble à la mémoire DRAM à ceci près que le condensateur utilise un matériau ferroélectrique pour retenir les données. Il s'agit généralement d'un film en céramique (PZT[26] ou PZTN[27]). Le principe de fonctionnement est le même que pour la cellule DRAM mais la mémoire FeRAM est non volatile du fait que son condensateur n'a pas de fuite. Elle ne consomme donc de l'énergie que lors des phases de lecture et d'écriture des données. La figure 1.11 illustre le stockage dans le condensateur dû à un cycle d'hystérésis de la charge en fonction de la polarisation appliquée, une polarisation négative traduisant un "1" et une polarisation positive correspondant à un niveau "0".

Pour lire l'état de la cellule, le transistor force le condensateur C_{FE} à prendre l'état "0". Ainsi, si le condensateur contenait déjà un "0", rien ne se passe, tandis que si un état "1" était stocké, les atomes vont se réorienter ce qui créera une pulsation de courant, détectable par un circuit de lecture.

[26] Titano-Zirconiate de Plomb
[27] Titano-Zirconiate de Plomb et Niobium

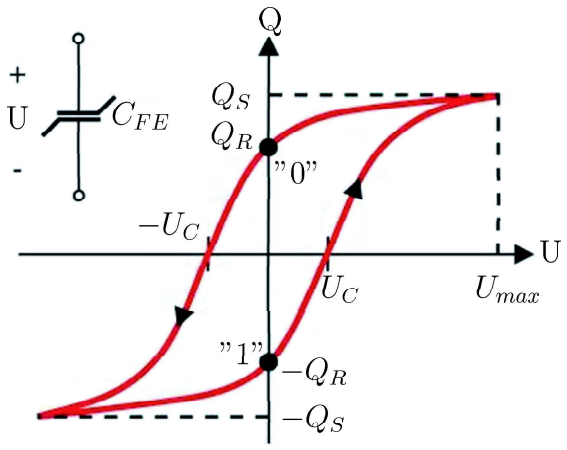


FIG. 1.11 – Principe de stockage de l'information dans le condensateur d'une cellule mémoire FeRAM.

Le tableau 1.3 résume les principales caractéristiques des mémoires FeRAM.

Mémoire FeRAM	
Taille de cellule (F^2)	30
Endurance en écriture/lecture	$10^{15}/10^{15}$
Temps de lecture (aléatoire)	$40ns + 80ns$ = lecture + ré-écriture car lecture destructive
Temps de programmation (/octet)	$80ns$
Temps d'effacement (/octet)	$80ns$
Scalabilité	Faible
Limite de scalabilité	Taille de la capacité
Possibilité de multi-niveaux	Non
Coût relatif par bit	Elevé
Maturité	Moyenne

TAB. 1.3 – Résumé des caractéristiques des mémoires FeRAM. [ITRS'05a]

1.1.4.3 Les mémoires à changement de phase

Les mémoires à changement de phase, appelées PCM[28], PRAM[29], CRAM[30] ou OUM[31] dans la littérature, reposent sur une modification de la résistivité d'un matériau

[28] Phase-Change Memory
[29] Phase-change Random Access Memory
[30] Chalcogenide Random Access Memory
[31] Ovonic Unified Memory, d'après le nom de son inventeur Stanford Ovshinsky

selon la phase dans laquelle celui-ci se trouve. La structure mémoire est représentée par la figure 1.12. Le matériau utilisé est un verre de chalcogénure qui bascule d'un état cristallin de faible résistivité à un état amorphe de forte résistivité. Les chalcogénures sont des composés chimiques comprenant un chalcogène, atome de la colonne 16 sur la table périodique des éléments et présentant un déficit de deux électrons (oxygène, soufre, sélénium, tellure ou polonium), en tant qu'ion négatif. Le chalcogénure le plus répandu dans les mémoires PCM utilise le tellure en alliage ternaire avec le germanium (Ge) et l'antimoine (Sb) selon $Ge_2Sb_2Te_5$, plus communément appelé GST [Pellizer'04]. Le changement d'état se fait par le passage d'un courant électrique qui provoque un échauffement par effet Joule. Le temps de basculement vers l'état amorphe est de l'ordre de quelques dizaines de nanosecondes.

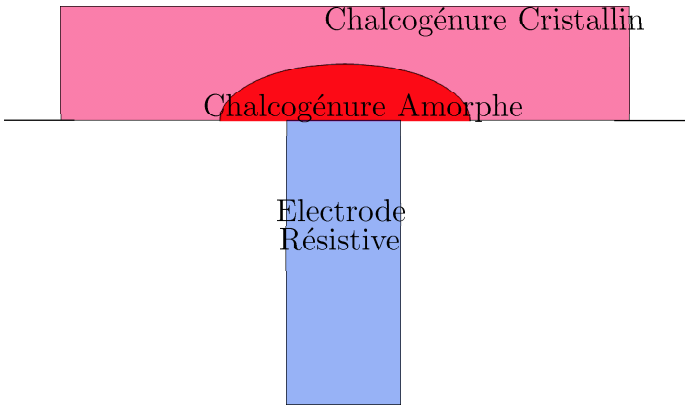


FIG. 1.12 – Cellule mémoire à changement de phase.

Le tableau 1.4 résume les principales caractéristiques des mémoires PCM.

Mémoire PCM	
Taille de cellule (F^2)	$10-16$
Endurance en écriture/lecture	$10^{12}/\infty$
Temps de lecture (aléatoire)	$60ns$
Temps de programmation (/octet)	$10ns$
Temps d'effacement (/octet)	$150ns$
Scalabilité	Bonne
Limite de scalabilité	Lithographie
Possibilité de multi-niveaux	Oui
Coût relatif par bit	Moyen
Maturité	Faible

TAB. 1.4 – Résumé des caractéristiques des mémoires PCM. [ITRS'05a]

1.1.4.4 Les mémoires à nano-cristaux

Dans les mémoires à nano-cristaux, la grille flottante que l'on trouve notamment dans une mémoire Flash est remplacée par une couche de nano-cristaux, généralement des nano-cristaux de silicium (Si) ou de germanium (Ge), qui servent de pièges discrets de charges [Tiwari'96], comme cela est le cas sur la figure 1.13.

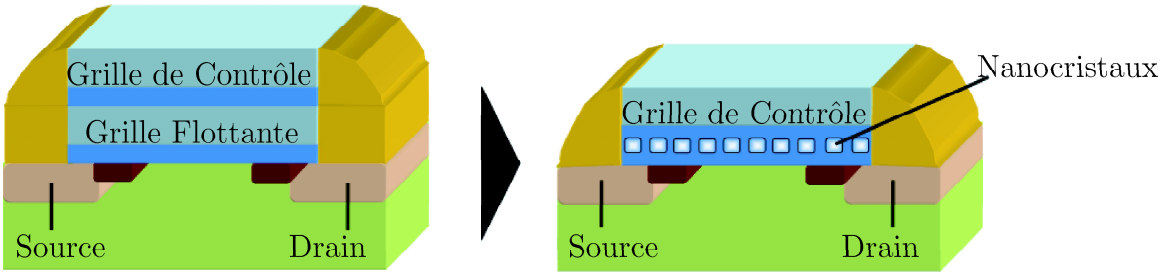


FIG. 1.13 – Cellule mémoire à nano-cristaux.

Le tableau 1.5 résume les principales caractéristiques des mémoires à nano-cristaux.

Mémoire à nano-cristaux	
Taille de cellule (F^2)	$4-10$
Endurance en écriture/lecture	$10^6/\infty$
Temps de lecture (aléatoire)	$60ns$
Temps de programmation (/octet)	$10\mu s$
Temps d'effacement (/octet)	$2ms/bloc$
Scalabilité	Correcte
Limite de scalabilité	Oxyde Tunnel, hautes tensions
Possibilité de multi-niveaux	oui
Coût relatif par bit	Moyen-Faible
Maturité	Bonne

TAB. 1.5 – Résumé des caractéristiques des mémoires à nano-cristaux [ITRS'05a] [ITRS'05b].

1.1.4.5 NRAM^{TM}

La mémoire NRAM^{TM} utilise les avancées de la recherche dans le domaine des nanotubes de carbone depuis 1991 pour réaliser une nouvelle cellule mémoire [Rueckes'00]. Chaque cellule est composée de plusieurs nanotubes suspendus à 13nm au-dessus d'une électrode. Une minuscule goutte d'or est déposée aux extrémités des tubes en guise de contact électrique supérieur. Une seconde électrode est située en-dessous de la première de sorte que lorsqu'un courant circule entre les deux électrodes,

les nanotubes sont attirés vers l'électrode inférieure et entrent en contact avec celle-ci. Si aucun courant n'est appliqué entre les deux électrodes, les nanotubes restent suspendus.

Pour savoir si les nanotubes touchent ou non l'électrode inférieure, on applique une tension sur l'électrode supérieure et on mesure le courant sur l'électrode inférieure ; si le courant passe, on lit un niveau "1", sinon on lit un niveau "0". La non-volatilité de cette mémoire vient du fait que les deux états de positionnement des nanotubes sont des états stables car la pression mécanique est compensée par les forces de Van der Waals. En effet, la figure 1.14 montre les deux états stables des nanotubes de carbone, qui sont des minima locaux de l'énergie totale, somme de l'énergie de la contrainte mécanique et de l'énergie due aux forces de Van der Waals entre les nanotubes et l'électrode.

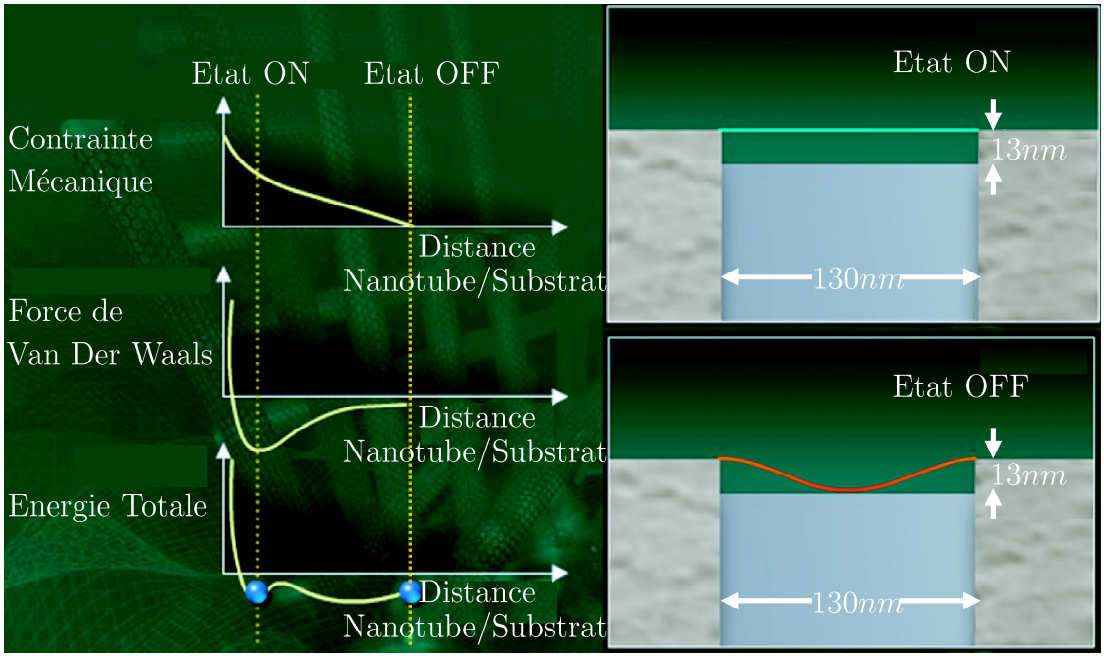


FIG. 1.14 – Cellule mémoire NRAM^{TM} à nanotubes de carbone. (Source : Nantero Inc.)

Le tableau 1.6 résume les quelques caractéristiques des mémoires NRAM^{TM} publiées à l'heure actuelle.

Mémoire NRAM^{TM}	
Endurance en écriture/lecture	$> 10^6$
Scalabilité	Bonne
Limite de scalabilité	Longueur du nanotube de carbone
Maturité	Moyenne

TAB. 1.6 – Résumé des caractéristiques des mémoires NRAM^{TM}.

A toutes ces nouvelles cellules mémoires s'ajoute une multitude d'autres axes de recherche basés autour de nouveaux matériaux, de polymères ou de molécules qui auraient la propriété de posséder au moins deux états stables, correspondant chacun à un niveau logique.

1.1.5 Marché des mémoires à semiconducteurs

La figure 1.15 présente le marché des mémoires volatiles de type SRAM et DRAM qui, au cours de ces dix dernières années, est resté quasi-constant, avec néanmoins des variations selon l'évolution globale, que l'on sait très cyclique, du marché des semiconducteurs. A noter notamment le pic de marché en l'an 2000 puis l'effondrement l'année suivante, suivi d'une reprise et d'un retour au niveau initial entre 2002 et 2007. Pendant ce temps, le marché des mémoires non-volatiles de type Flash n'a cessé de progresser, gagnant un facteur 10 en 10 ans.

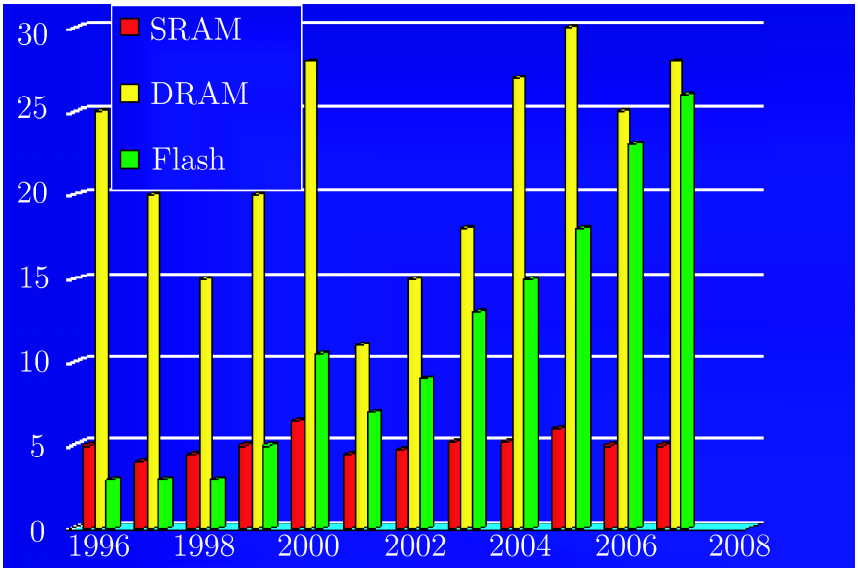


FIG. 1.15 – Evolution des marchés des mémoires SRAM, DRAM et Flash depuis 1996. (en Milliards de dollars)

Cette hausse considérable du marché des mémoires Flash est due au développement de nouvelles applications telles que la téléphonie mobile dont la plupart des appareils comportent une mémoire Flash de stockage interne, la photographie numérique avec l'essor des cartes mémoires ainsi que plus récemment le développement de mini-PCs où une mémoire Flash remplace les disques durs[32]. La figure 1.16 montre cette évolution du marché par application concernée.

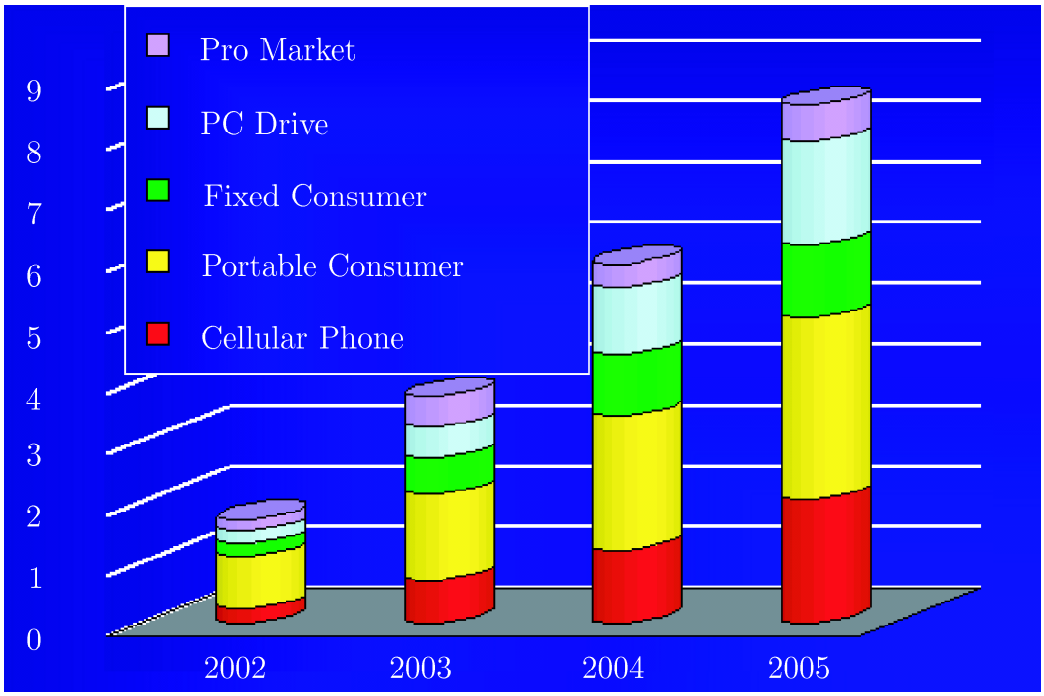


FIG. 1.16 – Evolution des marchés des mémoires non-volatiles par type d'application depuis 2002. (en Milliards de dollars)

La figure 1.17 montre quant à elle la comparaison entre l'évolution des mémoires de travail et l'évolution des mémoires de stockage dont les mémoires Flash font partie. Ainsi, lorsque l'on passe d'une génération à la suivante, tandis que la capacité de la mémoire de travail a augmenté de 10%, la capacité de la mémoire de stockage a augmenté d'un facteur 10, ce qui montre le potentiel gigantesque de ces mémoires. En effet, le développement actuel de nouvelles applications, telles que le remplacement des disques durs par des mémoires de type Flash dans les ordinateurs ultra-portables et portables, permet une augmentation considérable du marché des mémoires non-volatiles.

[32]les capacités de stockage sont certes assez limitées par rapport aux disques durs classiques mais offrent un gain important en place

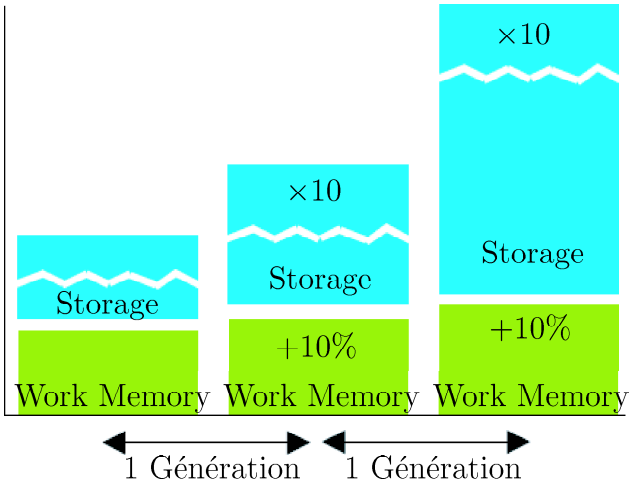


FIG. 1.17 – Prévisions de l'évolution des mémoires de travail et de stockage.

1.2 Les mémoires Flash

La cellule mémoire Flash repose sur le principe des technologies à grille flottante décrites précédemment en 1.1.2.1.

1.2.1 Structure et Principe de la cellule mémoire Flash

La cellule mémoire de type Flash est constituée d'un transistor à grille flottante appelé "transistor d'état" qui peut être représenté comme un transistor classique dont la grille serait mise en série avec un condensateur C_{pp}. Cette grille que l'on appellera alors "grille flottante" (FG[33]) peut alors stocker une charge et la seconde électrode du condensateur devient la "grille de contrôle" (CG[34]) de la cellule, dont le schéma électrique est représenté en figure 1.18. La transparence de la barrière, électriquement schématisée par une source de courant I permet l'injection de charges, stockées dans la grille flottante et qui modifient la tension de seuil V_T du transistor MOS selon la relation :

$$V_T \approx V_{T0} - \frac{Q_{FG}}{C_{pp}} \tag{1.1}$$

où V_{T0} est la tension de seuil naturelle de la cellule, Q_{FG} la quantité de charges dans la grille flottante et $C_{pp} = \frac{\varepsilon_{ox}.\varepsilon_0.S_{pp}}{t_{pp}}$ la capacité interpoly où ε_{ox} est la permittivité diélectrique relative de l'oxyde utilisé, ε_0 est la permittivité diélectrique du vide, S_{pp} la surface et t_{pp} l'épaisseur de l'isolant séparant les deux grilles C_{pp}.

[33]Floating Gate
[34]Control Gate

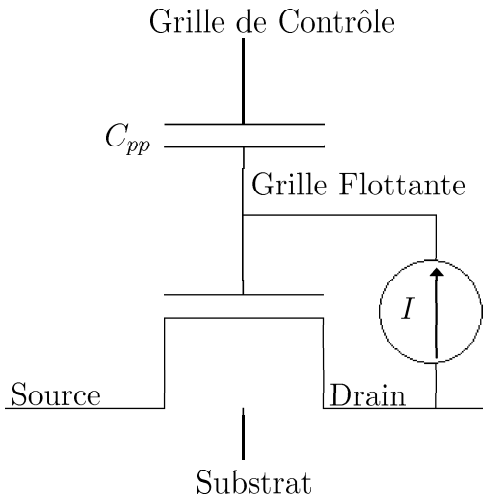


FIG. 1.18 – Schéma de principe du transistor d'état.

Ainsi, d'après l'équation (1.1), une charge positive diminue la tension de seuil du transistor tandis qu'une charge négative augmente cette tension de seuil, ce qui permet de définir deux états de part et d'autre de la tension de seuil naturelle V_{T_0}, visibles sur la figure 1.19 :

- un état programmé au niveau haut ($V_T = V_{T_{Haut}}$) correspondant au "0" logique ;
- un état effacé au niveau bas ($V_T = V_{T_{Bas}}$) correspondant au "1" logique.

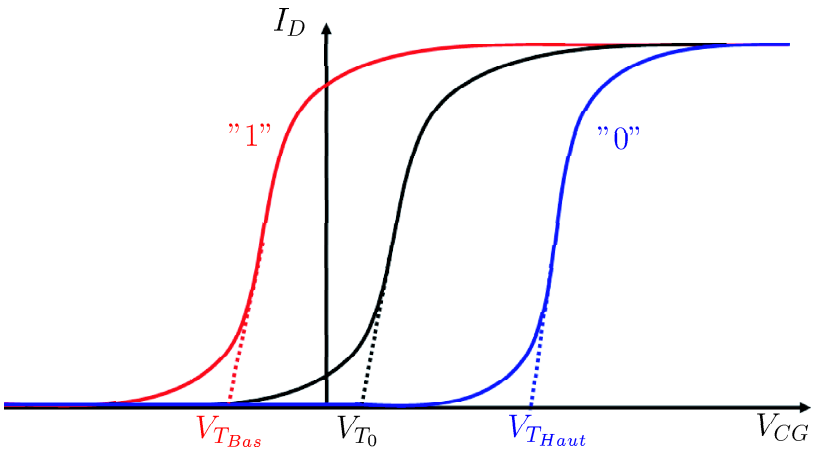


FIG. 1.19 – Caractéristiques de la cellule *Flash*.

On définit la fenêtre de programmation comme étant la différence entre les tensions de seuil des deux niveaux, soit $V_{T_{Haut}} - V_{T_{Bas}}$.

1.2.2 Fonctionnement de la cellule mémoire Flash

1.2.2.1 Lecture de la cellule mémoire Flash

En se basant sur la figure 1.20, nous voyons qu'il est tout à fait possible de distinguer ces deux états en mesurant par exemple le courant I_D pour une polarisation $V_{CG} = 0V$.

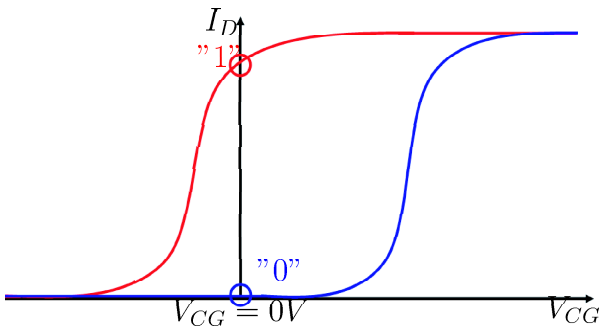


FIG. 1.20 – Caractéristiques de lecture de la cellule *Flash*.

Ainsi, à $V_{CG} = 0V$, soit le courant mesuré est élevé (c'est-à-dire supérieur à une valeur de courant de référence fixée) et on mesure un état "1", soit le courant mesuré est faible (c'est-à-dire inférieur à une valeur de courant de référence fixée) et on mesure un état "0".

1.2.2.2 Programmation de la cellule mémoire Flash

Pour programmer la cellule, il faut charger négativement la grille flottante ($Q_{FG} < 0$). Deux mécanismes peuvent pour cela être utilisés :

Injection de porteurs chauds

L'injection de ces électrons chauds, ainsi appelés en raison de leur vitesse élevée (ce qui équivaut à une énergie cinétique élevée), a lieu du côté du drain. La composante du champ électrique le long du canal produit une distribution d'électrons chauds qui sont injectés dans la grille flottante par la composante transverse du champ électrique. L'injection est effective si le champ longitudinal est suffisamment élevé (on est en saturation) et si le champ transverse est orienté de la grille flottante vers le drain. Le champ longitudinal est créé en polarisant le drain à environ $+4V$ avec la source à la masse, tandis que le champ transverse est obtenu par une polarisation positive d'environ $+8V$ sur la grille de contrôle. Ce phénomène d'injection d'électrons chauds, décrit dans la figure 1.21, peut être favorisé par une polarisation légèrement négative du substrat : c'est ce que l'on appelle le CHISEL[35]. En polarisant négativement le

[35]CHannel Initiated Secondary ELectron

substrat, on augmente l'efficacité d'injection en créant une génération supplémentaire de porteurs chauds grâce à une seconde ionisation par impact.

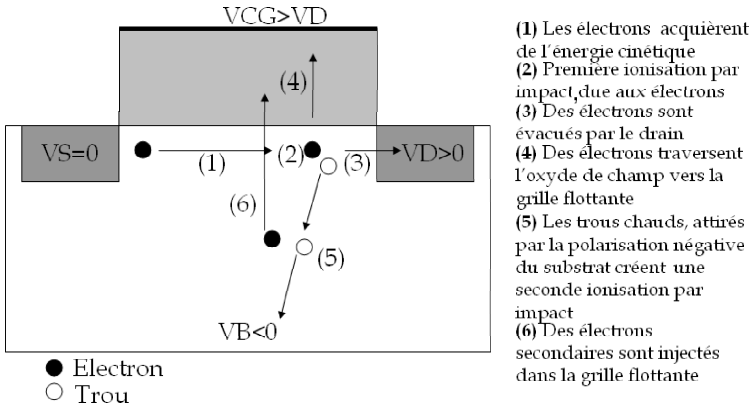


FIG. 1.21 – Principe complet de programmation de la cellule par porteurs chauds.

Effet Fowler-Nordheim

L'effet Fowler-Nordheim est un effet tunnel qui a lieu lorsque l'on amincit la largeur de la barrière par une polarisation aux bornes de l'oxyde. La barrière de potentiel passe alors d'une forme trapézoïdale, classique du courant tunnel direct, à une forme quasi-triangulaire, ce que montre la figure 1.22 [FowlerNordheim'28]. Le courant tunnel Fowler-Nordheim est détaillé au chapitre 2.

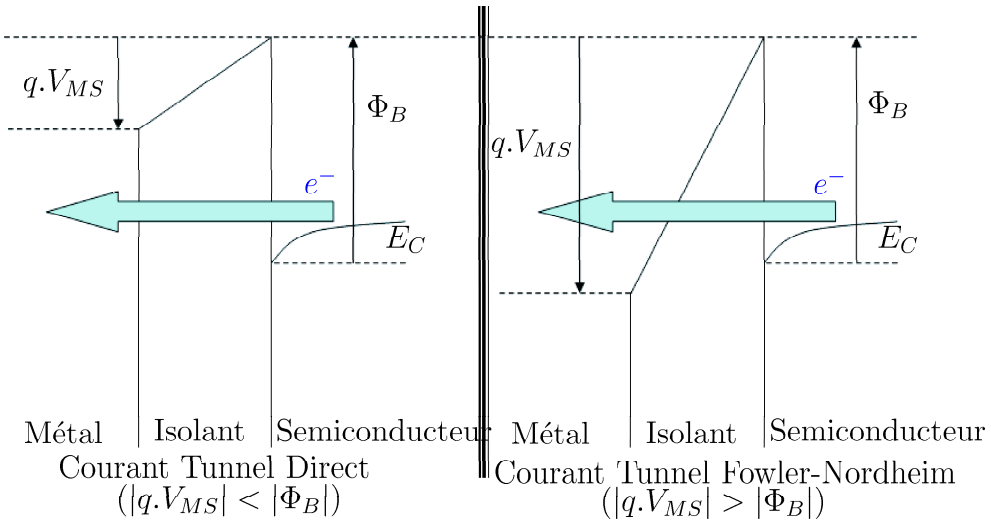


FIG. 1.22 – Principe de programmation par effet Fowler-Nordheim.

1.2.2.3 Effacement de la cellule mémoire Flash

Afin d'effacer la cellule, il faut évacuer les charges négatives de la grille flottante. On utilise pour cela l'effet Fowler-Nordheim décrit précédemment en mettant la grille de contrôle à la masse et en polarisant positivement à une tension suffisante la source, le drain et le substrat. Le potentiel de grille flottante diminue alors par couplage capacitif entre la grille de contrôle et la grille flottante ce qui crée un champ électrique dans l'oxyde tunnel. On obtient par conséquent un courant Fowler-Nordheim de la grille flottante vers le drain. On a donc l'équivalent d'une injection de charges positives dans la grille flottante ($Q_{FG} > 0$).

1.2.3 Architectures des matrices de mémoires Flash

Il existe de nombreuses façons d'organiser les cellules mémoires entre elles en vue de créer une matrice mémoire de taille plus ou moins grande, c'est ce que l'on appelle les différentes architectures. Il existe deux principales architectures : les architectures "NOR", développée par Intel et "NAND", développée par Toshiba, auxquelles s'ajoutent nombre d'architectures plus marginales telles que "And" d'Hitachi, "Dinor"[36] de Mitsubishi ou encore "T-Poly" de Sandisk. Nous ne détaillerons dans ce manuscrit que les deux architectures principales qui sont de loin les plus utilisées.

1.2.3.1 Architecture NOR

Dans une matrice mémoire en architecture NOR, l'ensemble des cellules sont connectées en parallèle entre les lignes de Source et de Bit avec un accès individuel à chaque cellule, comme décrit en figure 1.23.

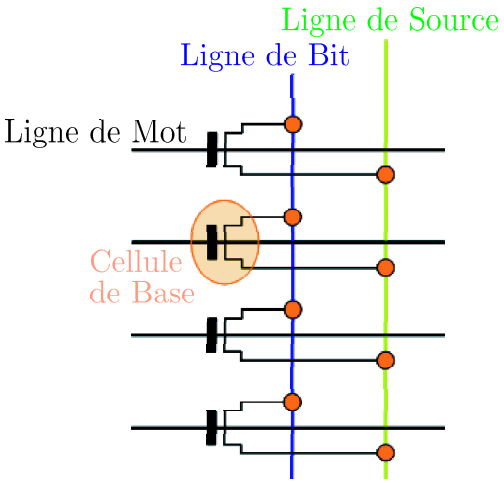


FIG. 1.23 – Schéma des connections entre cellules en architecture NOR.

[36] Divided bit-line nor

Nous pouvons résumer les principales caractéristiques des mémoires Flash en architecture NOR dans le tableau 1.7.

Mémoire Flash NOR	
Taille de cellule (F^2)	10
Endurance en écriture/lecture	$10^6/\infty$
Temps de lecture (aléatoire)	$60ns$
Temps de programmation (/octet)	$1\mu s$
Temps d'effacement (/octet)	$1s/secteur$
Scalabilité	Correcte
Limite de scalabilité	Oxyde Tunnel, hautes tensions
Possibilité de multi-niveaux	Oui
Coût relatif par bit	Moyen
Maturité	Très bonne

TAB. 1.7 – Résumé des caractéristiques des mémoires Flash NOR [ITRS'05a] [ITRS'05b].

1.2.3.2 Principe de fonctionnement de la mémoire Flash en architecture NOR

L'architecture NOR permettant de disposer d'un contact individuel pour la Grille de Contrôle, la Source, le Drain et d'un contact commun pour le Substrat, nous pouvons directement utiliser la description faite en 1.2.2 du fonctionnement général d'une mémoire Flash en appliquant chacune des polarisations indiquées.

1.2.3.3 Architecture NAND

Les cellules sont placées en série entre la ligne de Source et la ligne de Bit par groupe de 4, 8, 16 voire 32 cellules. Pour accéder à une cellule, il faut « passer » à travers toutes les cellules voisines. On peut également noter la présence de transistors de sélection en bouts de chaîne pour accéder à la ligne de Source ainsi qu'à la ligne de Bit, comme décrit en figure 1.24.

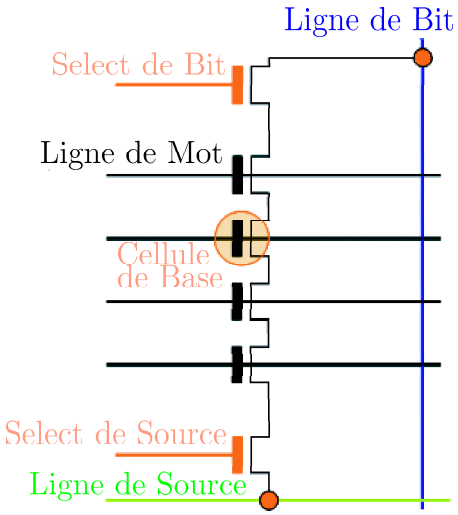


FIG. 1.24 – Schéma des connections entre cellules en architecture NAND.

Nous pouvons résumer les principales caractéristiques des mémoires Flash en architecture NAND dans le tableau 1.8.

Mémoire Flash NAND	
Taille de cellule (F^2)	4
Endurance en écriture/lecture	$10^6/\infty$
Temps de lecture (aléatoire)	$60ns$ en série
Temps de programmation (/octet)	$200\mu s/page$
Temps d'effacement (/octet)	$2ms/bloc$
Scalabilité	Correcte
Limite de scalabilité	Oxyde Tunnel, hautes tensions
Possibilité de multi-niveaux	Oui
Coût relatif par bit	Faible
Maturité	Très bonne

TAB. 1.8 – Résumé des caractéristiques des mémoires Flash NAND [ITRS'05a] [ITRS'05b].

Comme le montre la figure 1.25, l'architecture NAND présente une bien meilleure intégration et offre donc des possibilités importantes de réduction des coûts de production. En effet, la suppression des contacts individuels de Drain et de Source qui existent dans l'architecture NOR permet une réduction considérable de l'espacement entre les cellules. Si dans les deux cas, nous pouvons mesurer sur le layout une dimension $2F$ perpendiculairement au canal, l'espace nécessaire au passage de via métalliques pour venir contacter les zones d'implants Source et Drain impose

une dimension $5F$ pour l'architecture NOR contre $2F$ pour l'architecture NAND parallèlement au canal.

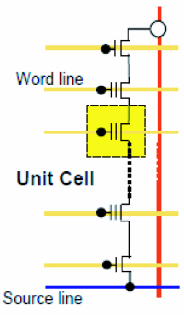
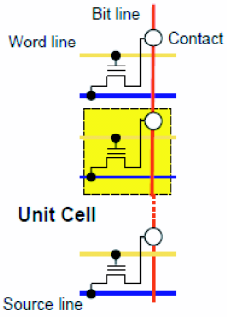
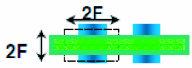
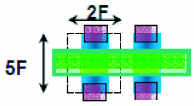
	NAND	NOR
Matrice		
Layout		
Coupe		
Taille	$4F^2$	$10F^2$

FIG. 1.25 – Tableau récapitulatif des architectures NOR et NAND.

1.2.3.4 Principe de fonctionnement de la mémoire Flash en architecture NAND

La cellule mémoire Flash élémentaire en architecture NAND fonctionne sur le principe décrit dans le paragraphe 1.2.1. La principale différence réside dans le fait qu'il faille polariser convenablement toutes les cellules voisines pour utiliser correctement la cellule adressée, polarisations que nous allons détailler dans la suite de ce paragraphe.

D'après la figure 1.26, la lecture se fait en polarisant la grille de contrôle de la cellule que l'on veut lire, à une tension fixée et déterminée à l'avance, par exemple 0V. Il faut alors mesurer le courant de drain de cette cellule. Pour cela, il faut bien évidemment que toute la chaîne de cellules soit passante. Ne connaissant pas l'état des cellules voisines de la cellule à lire, il faut polariser leurs grilles de contrôle respectives à une polarisation supérieure à leur tension de seuil, polarisation que nous noterons V_{pass}. Les deux transistors de sélection doivent également être rendus passants, ce qui est réalisé grâce à cette même polarisation V_{pass}, pour avoir accès aux lignes de Source et de Bit Line (Drain). La source et le substrat sont alors maintenus à 0V, tandis que la Bit Line est polarisée à une tension de l'ordre de $1V$.

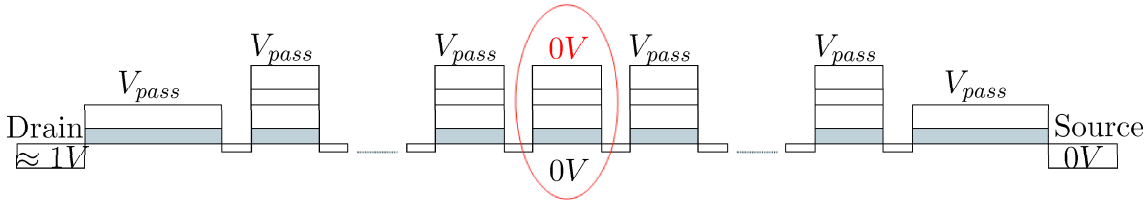


FIG. 1.26 – Polarisations typiques en lecture d'une cellule Flash NAND.

Les mémoires Flash en architecture NAND utilisent le mécanisme Fowler-Nordheim, aussi bien en programmation qu'en effacement, ce sont donc des mémoires à grille flottante de type FLOTOX. Ces cellules sont dites "Full-Fowler".

Pour programmer une cellule de la chaîne, phase représentée en figure 1.27, il faut que sa source soit flottante, que son drain et le substrat soient à $0V$, et que sa grille de contrôle reçoive une forte polarisation V_{prog} de l'ordre de $16V$. Nous pouvons donc appliquer directement sur la grille de contrôle le signal de programmation, mais une fois de plus pour amener les polarisations désirées sur le drain et la source de la cellule, il faut rendre toutes les cellules ainsi que le transistor de sélection de la Bit Line passants grâce à la tension V_{pass}. En revanche pour que la source reste flottante, le transistor de sélection de la ligne de source doit être bloqué par une polarisation V_{GS} négative.

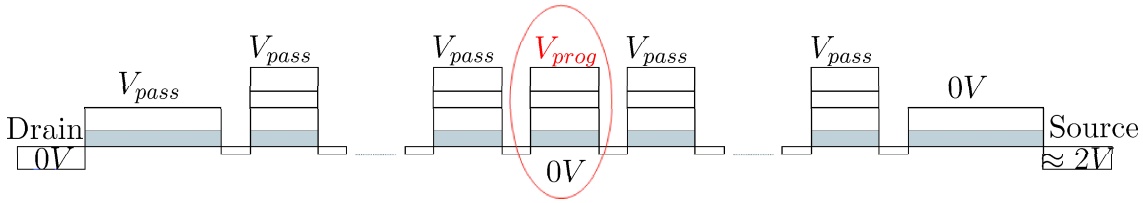


FIG. 1.27 – Polarisations typiques en programmation d'une cellule Flash Nand.

En effacement, il n'est pas possible d'effacer une seule cellule de la chaîne. L'effacement se fait pour l'ensemble de la chaîne en polarisant toutes les grilles de contrôle à $0V$. On efface alors en polarisant à une forte tension positive V_{eff} de l'ordre de $16V$, le substrat, la ligne de Bit, la ligne de Source ainsi que les deux grilles des transistors de sélection BSL et GSL. La figure 1.28 résume l'ensemble de ces conditions de polarisation.

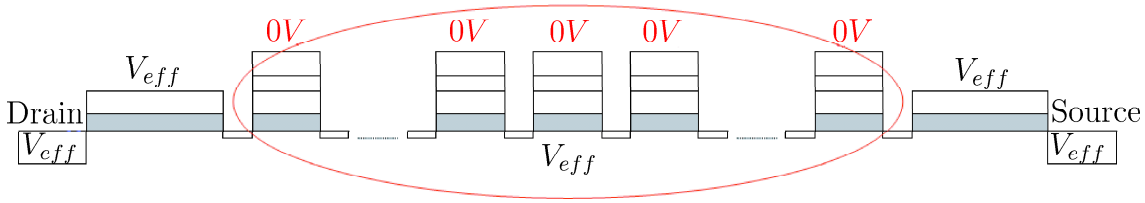


FIG. 1.28 – Polarisations typiques en effacement d'une cellule Flash Nand.

1.2.3.5 Marché des mémoires Flash NOR et NAND

En l'an 2000, le marché des mémoires Flash était largement dominé par l'architecture NOR qui représentait environ 80% du marché contre seulement 20% pour l'architecture NAND. Après les baisses de marché en 2001 et 2002, le volume global des mémoires Flash a augmenté avec une part de plus en plus importante prise par l'architecture NAND jusqu'à une inversion des forces entre 2004 et 2005. En 2007, l'architecture NAND représentait environ 60% du marché total des mémoires Flash. La figure 1.29 résume l'évolution du marché des mémoires Flash en architecture NOR et NAND depuis 2000.

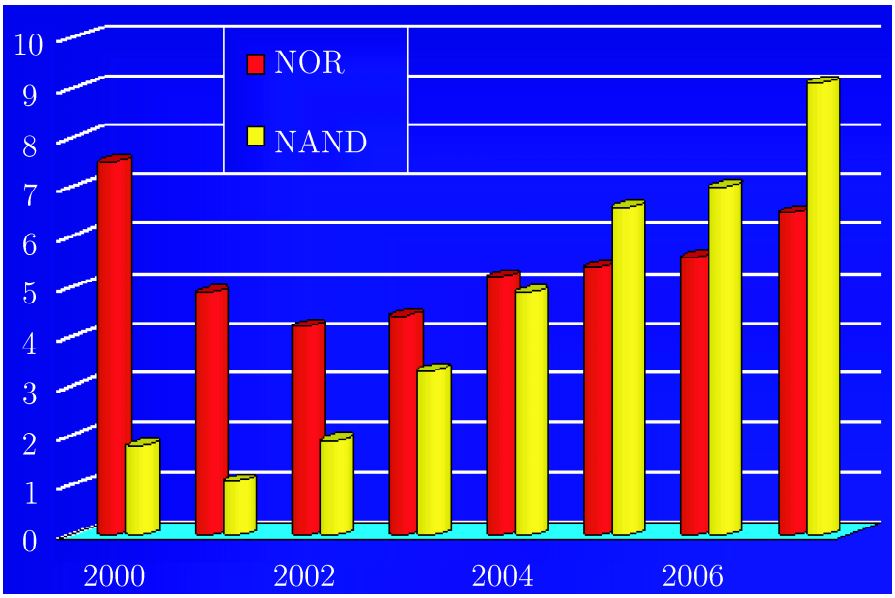


FIG. 1.29 – Evolution des marchés des mémoires Flash en Architectures NOR et NAND depuis 2000. (en Milliards de dollars)

1.2.4 Perturbations intervenant au cours du fonctionnement

A l'intérieur d'une matrice de cellules mémoires, représentée en figure 1.30, au cours d'une opération qui adresse une des cellules, des perturbations, appelées disturb, peuvent apparaître sur les cellules voisines.

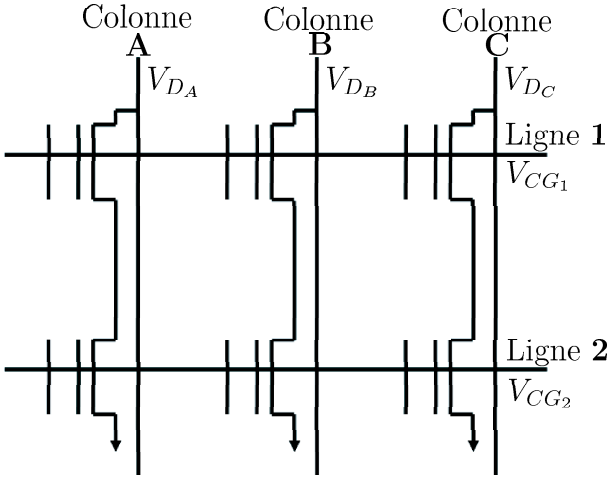


FIG. 1.30 – Schéma d'une matrice mémoire.

Une colonne représente une **B**it**L**ine (ligne de drain) et une ligne représente une **W**ord**L**ine (ligne de grille de contrôle).

A partir de la matrice de la figure 1.30, si l'on considère la cellule adressée **B2**, alors les cellules **B1** et **C2** peuvent subir des perturbations, selon leur état et l'opération effectuée sur la cellule **B2**.

Dans la suite de ce manuscrit nous utiliserons couramment le terme de "disturb" plutôt que "perturbation" du fait de son utilisation quasi-systématique dans le milieu industriel. Dans les chapitres suivants, toutes les perturbations étudiées seront des perturbations dues à la tension de grille Vprog en programmation, c'est pourquoi nous nous contenterons de les désigner par le terme "perturbation" ou "disturb".

Il existe 3 principaux types de disturb dans le cas d'une matrice NOR :
- program disturb ;
- drain disturb ;
- read disturb.

et 3 principaux types de disturb dans le cas d'une matrice NAND :
- program disturb ;
- pass disturb ;
- read disturb.

1.2.4.1 Program Disturb ou perturbation due à la tension de grille Vprog en programmation (architectures NOR et NAND)

Ce type de perturbation intervient si la cellule **C2** est non-programmée ou à l'état effacé et que la cellule **B2** est en phase de programmation. La cellule **C2** a peu d'électrons sur sa grille flottante et possède un V_T faible. Quand on applique +15V sur la WL, le champ électrique à travers l'oxyde de champ peut être suffisant pour créer une injection d'électrons du substrat vers la grille flottante, soit une augmentation du V_T. Dans certains cas, la cellule **C2** peut être légèrement programmée accidentellement : c'est le "soft-write".

1.2.4.2 Pass Disturb ou perturbation due à la tension de passage Vpass en programmation (architecture NAND)

Il se produit si la cellule **B1** est à l'état effacé et que la cellule **B2** est en phase de programmation. Afin de pouvoir programmer la cellule **B2**, il faut rendre la cellule **B1** passante ce qui se fait en appliquant une polarisation positive de l'ordre de 8V sur la WL. Ainsi se crée un champ électrique à travers l'oxyde qui peut provoquer une injection d'électrons dans la grille flottante et l'on a également risque de "soft-write".

1.2.4.3 Drain disturb ou perturbation due à la tension de drain Vdrain en programmation (architecture NOR)

Il se produit si la cellule **B1** est à l'état programmé et que la cellule **B2** est en phase d'effacement. Il se crée alors un champ électrique entre la grille flottante et le drain. Des électrons peuvent alors être évacués de la grille flottante vers le drain, causant une baisse de V_T et un léger effacement.

1.2.4.4 Read Disturb ou perturbation due à la lecture (architectures NOR et NAND)

Ce mécanisme de perturbation se produit si la cellule **C2** est à l'état effacé et que la cellule **B2** est lue [Ielmini'01]. La WL commune reçoit une tension positive de quelques Volts, tandis que la BL de la cellule lue est à environ +1V. Les cellules non lues reçoivent quant à elles 0V sur leur source, BL et substrat. Cette fois encore un "soft-write" est possible.

1.3 Conclusion

Ce premier chapitre nous a permis de situer les mémoires Flash à l'intérieur de la vaste famille des mémoires à semiconducteurs, allant des mémoires développées de longue date aux mémoires émergentes au centre des recherches actuelles. Nous donnons ensuite une évolution du marché des différentes mémoires au cours des dix dernières années qui a vu l'essor des mémoires Flash, notamment en architecture

NAND. Nous avons également pu présenter les différentes architectures de matrices mémoires et en décrire le fonctionnement spécifique. Nous terminons le chapitre par une description des phénomènes de perturbations entre cellules à l'intérieur d'une matrice mémoire.

Références bibliographiques du chapitre 1

[FowlerNordheim'28] R.H. Fowler, L. Nordheim
"Electron Emission in intense electric fields"
Proc. Soc. London Ser., A119, 781, pp.173, 1928.

[Chen'77] P. C. Chen
"Threshold-alterable Si-gate MOS devices"
IEEE Trans. Elec. Devices, vol. ED-24, pp.584, 1977.

[Tiwari'96] S. Tiwari, F. Rana , K. Chan, W. Chen
"Single Charge and Confinement Effects in Nano-Crystal Memories"
Appl. Phys. Lett., Vol. 69, pp.1232, 1996.

[Pavan'97] P. Pavan, R. Bez, P. Olivo, E. Zanoni
"Flash Memory Cells - An overview"
Proceedings of the IEEE, Vol. 85, No. 8, pp.1248, 1997.

[Brown'98] W.D. Brown, J.E. Brewer
"Nonvolatile semiconductor Memory Technology"
IEEE Press, New York, 1998.

[Rueckes'00] T. Rueckes, K. Kim, E. Joselevich, G.Y. Tseng, C.L. Cheung, C.M. Lieber
"Carbon Nanotube-Based NonVolatile Random Access Memory for Molecular Computing"
Science, Vol. 317, Issue 5476, pp.94, 2000.

[Ielmini'01] D. Ielmini, A.S. Spinelli, A.L. Lacaita, L. Confalonieri, A. Visconti
"New technique for fast characterization of SILC distribution in Flash arrays"
Proceedings of IRPS, pp.73-80, 2001.

[Pellizer'04] F. Pellizer et al.
"Novel μTrench Phase-Change Memory cell for embedded and stand-alone Non-Volatile Memory applications"
Proceedings of VLSI Technology, pp.18-19, 2004.

[ITRS'05a] International Technology Roadmap for Semiconductors
"Emerging Research Devices"
ITRS Roadmap, 2005 Edition.

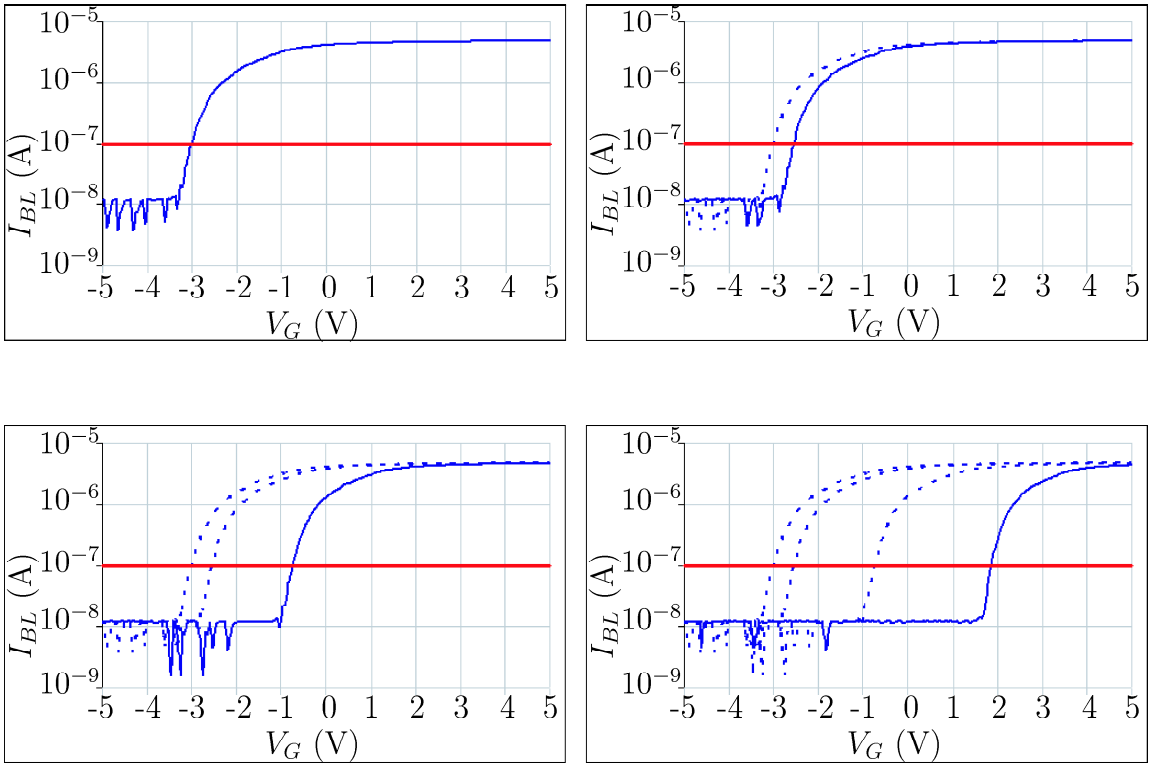
[ITRS'05b] International Technology Roadmap for Semiconductors
"Process Integration, Devices and Structures"
ITRS Roadmap, 2005 Edition.

[Pirovano'05] A. Pirovano, R. Bez
"Alternatives to Conventional Non-Volatile Memory : Status and Perspective"
Proceedings of ICMTD, pp.17, 2005.

[Lee'06] J.D. Lee, C.K. Lee, M.W. Lee, H.S. Kim, K.C. Park, W.S. Lee
"A new programming disturbance phenomenon in NAND Flash Memory by Source/Drain Hot-Electrons generated by GIDL current"
Proceedings of NVSMW, pp.31, 2006.

Chapitre 2

Etude des méthodes de programmation



Sommaire

Ce deuxième chapitre a pour objectif de présenter l'étude de diverses méthodes visant à améliorer la tenue de nos cellules aux tests en cyclage. Nous présenterons tout d'abord en quoi consiste ces tests en cyclage ainsi que leur intérêt, puis nous nous appuierons sur des méthodes décrites dans la littérature que nous appliquerons à nos cellules mémoires, dont nous détaillerons les structures au cours de ce chapitre. Nous traiterons dans un premier temps de l'utilisation de signaux de très courte durée, en vue de s'affranchir de la création de pièges dans l'oxyde, puis nous aborderons l'étude de la forme des signaux qui peut être optimisée pour minimiser le champ électrique à travers l'oxyde tunnel. Nous décrirons puis mettrons en œuvre un algorithme de programmation dit "intelligent" qui élimine la fermeture classique de fenêtre, garantissant ainsi la différenciation des deux niveaux logiques stockés par la cellule même après un nombre élevé de cycles de programmation/effacement.

2.1 Test en endurance ou Cyclage

Les mémoires non volatiles et en particulier les mémoires *Flash* sont conçues pour être programmées et effacées un grand nombre de fois au cours de leur vie (typiquement 10^5 voire 10^6 cycles pour une durée de vie d'environ 10 ans du produit). Ainsi, un test en endurance a été développé pour évaluer de façon accélérée leur tenue à cette répétition de phases d'écriture et d'effacement : c'est ce que l'on appelle un test en endurance ou cyclage. Nous pouvons immédiatement préciser que dans l'ensemble de ce manuscrit, les termes "programmation" et "effacement" représenterons respectivement une injection de charges négatives dans la grille flottante et leur retrait de cette grille flottante, selon les conventions usuelles pour les mémoires Flash.

2.1.1 Principe du test en endurance

Au cours d'un cyclage, la cellule est successivement programmée puis effacée en boucle, avec une lecture régulière des tensions de seuil du niveau programmé et du niveau effacé. L'évolution des tensions de seuil en fonction du nombre de cycles appliqués est représentée afin de mettre en évidence la variation de la fenêtre de programmation et de vérifier si les niveaux logiques restent distincts quel que soit le nombre de cycles subis par la cellule mémoire.

2.1.2 Théorie du test en endurance

Au fil de ces cycles de programmation/effacement, l'oxyde de champ va progressivement être dégradé avec la création de pièges. La génération des pièges est due aux champs électriques élevés dans l'oxyde lors des phases de programmation/effacement qui vont accélérer des électrons qui transitent à travers l'oxyde tunnel, créant des pièges dans l'oxyde, ces pièges pouvant être remplis par des charges [Euzent'81].

Au cours de ces cycles, la fenêtre de programmation va peu à peu se refermer du fait du piégeage des électrons lors de leur injection.

Lors des phases de programmation (ou d'effacement) sur une cellule vierge, le courant Fowler-Nordheim de densité J_{FN} est mis en jeu pour l'injection des charges à travers l'oxyde tunnel pour des polarisations positives (ou négatives) sur la grille de contrôle, selon les équations fournies par N. Baboux [Baboux'03].

$$I_{fg} = f_0(V_{fg}) = S_{tun}.J_{FN}(\frac{V_{fg} - K}{t_{tun}}) \tag{2.1}$$

avec V_{fg} la tension de la Grille Flottante, t_{tun} et S_{tun} respectivement l'épaisseur et la surface de l'oxyde tunnel et K un potentiel constant prenant en compte la variation de champ électrique aux bornes de l'oxyde tunnel, due aux charges piégées.

Pour une cellule avec un oxyde tunnel dégradé, la fonction d'injection devient :

$$I_{fg} = f(V_{fg}) = f_0(V_{fg} - \Delta V_{fg}) \tag{2.2}$$

avec pour $V_{fg} > 0$:

$$\Delta V_{fg} = \frac{t_{tun}}{\epsilon.S_{tun}}.Q_{ot}.(\bar{xr} - 1) \tag{2.3}$$

et pour $V_{fg} < 0$:

$$\Delta V_{fg} = \frac{t_{tun}}{\epsilon.S_{tun}}.Q_{ot}.\bar{xr} \tag{2.4}$$

où Q_{ot} est la quantité de charges piégées dans l'oxyde, ϵ est la permittivité totale de l'oxyde et $\bar{xr}$ la position relative moyenne de la distribution des charges dans l'oxyde, dirigée du substrat vers la grille flottante.

Les fonctions f_0 et f reliant la tension de la grille flottante au courant de programmation, respectivement avant et après dégradation de l'oxyde tunnel, sont tracées dans la figure 2.1

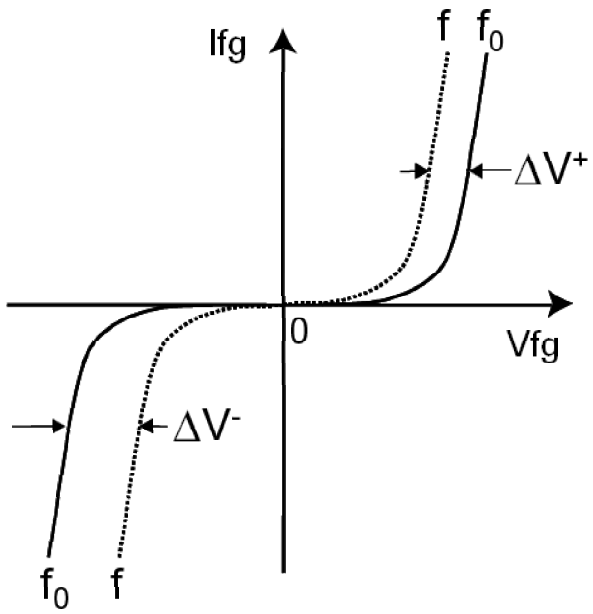


FIG. 2.1 – Courants d'injection sur cellule vierge (f0) et cellule cyclée (f)

La variation ΔV_{th} de tension de seuil des cellules, due à ces charges piégées dans l'oxyde tunnel, est donnée par :

$$\Delta V_{th} = \frac{-1}{\alpha_{fg}} \Delta V_{fg} \tag{2.5}$$

où α_{fg} est le coefficient de couplage entre la grille flottante et la grille de contrôle.

Ainsi, au fur et à mesure que des charges se piègent dans l'oxyde tunnel lors du cyclage de la cellule, l'efficacité d'injection diminue, le niveau effacé est un peu moins effacé alors que le niveau programmé est un peu moins programmé et la fenêtre de programmation se referme.

Cette dégradation des oxydes lors des opérations de cyclage peut également être vue comme un décalage de la tension de seuil naturelle V_{T_0} de la cellule avec le nombre de cycles [Tseng'06].

$$\bar{xr}.Q_{ot} = \alpha_{fg}.(V_{T_0} - V_{T_0,cycles})C_{tun}/q \tag{2.6}$$

où C_{tun} est la capacité de l'oxyde tunnel et q la charge élémentaire de l'électron.

En résumé, le test en endurance consiste donc à effectuer un grand nombre de cycles de programmation/effacement et à relever la fenêtre de programmation dont la fermeture traduit le vieillissement de l'oxyde de champ qui limite l'injection de porteurs par effet tunnel.

La figure 2.2 a) présente la fermeture de fenêtre observée classiquement, avec un rapprochement des deux niveaux pouvant compromettre leur différenciation et le bon fonctionnement de la cellule. La figure 2.2 b) présente quant à elle un type de cyclage où le niveau du V_T programmé augmente avec le nombre de cycles, ce qui aurait a priori un effet plutôt bénéfique d'écartement des niveaux mais on peut remarquer que le niveau effacé est beaucoup plus dégradé que sur la figure 2.2 a) [Lee'04]. Au final, ce type de dégradation pose autant de problèmes de séparation des états que la dégradation classique. En effet, plusieurs problèmes peuvent se poser du point de vue de la lecture des niveaux logiques :

1. si la lecture du niveau logique répond à un gabarit de tension de seuil, le fait que le niveau de V_T bas passe de $+2,5V$ à $+4.5V$ comme sur la figure 2.2 b) risque davantage de causer une sortie de ce gabarit qu'un passage de $+2,5V$ à $+4V$ comme sur la figure 2.2 a).
2. si dans la matrice de cellules, plusieurs niveaux de vieillissement cohabitent, en apparence la fenêtre de programmation reste bien ouverte mais il faut en réalité comparer le niveau de V_T bas à 10^5 cycles avec non pas le niveau de V_T haut à ce même nombre de cycles mais avec le niveau de V_T haut initial. Ainsi, la fenêtre de programmation devient beaucoup plus fermée qu'il n'y paraît.

On explique le phénomène de la figure 2.2 b) par la présence d'états d'interface qui sont créés par la rupture de liaisons Si-H due au champ électrique. En effet, le Silicium et le Dioxyde de Silicium étant respectivement cristalline et amorphe, cela crée de fortes distensions à l'interface ainsi que des liaisons pendantes. Des atomes d'hydrogène sont classiquement apportés lors du recuit final de la plaquette afin de passiver ces liaisons pendantes à l'interface Si/SiO2. Cependant, lorsque de forts champs électriques sont appliqués à travers l'oxyde tunnel, les liaisons Si-H, inactives

électriquement, de faible énergie de liaison, sont rompues et engendrent des états d'interface. Des études ont également montré que lorsque l'on crée un état d'interface, on crée en parallèle un piège dans l'oxyde [Jepson'77][Bénard'08].

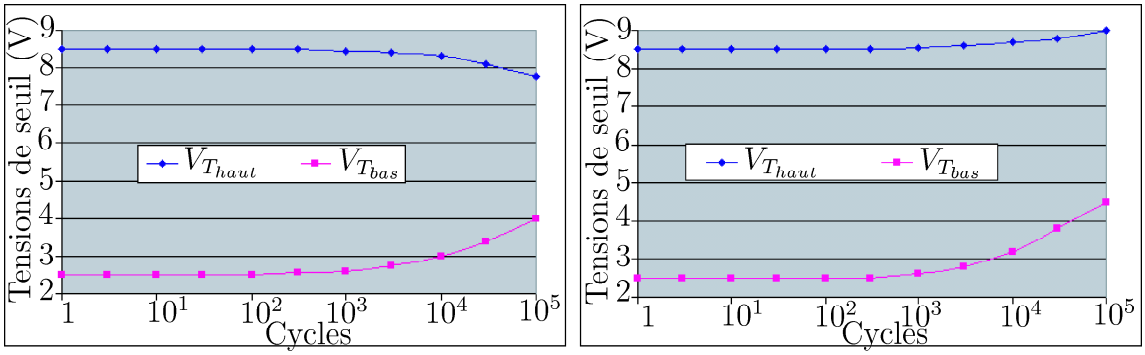


FIG. 2.2 – Variation des tensions de seuil lors d'un test en endurance a) "typique", b) avec présence d'états d'interface.

En vue d'améliorer la tenue des cellules à un nombre élevé de cycles de programmation/effacement, il est donc particulièrement important d'utiliser des oxydes de bonne qualité, présentant naturellement le moins de liaisons pendantes et/ou avec des liaisons Si-H de faible énergie ou Si-X, X étant un autre élément d'énergie de liaison plus élevée (Deutérium,...) limitant ainsi la création ultérieure d'états d'interface et de pièges dans le volume de l'oxyde [Denais'05].

Différentes voies peuvent également être explorées au niveau plus "software" afin de limiter la dégradation de l'oxyde tunnel, telles que des études sur la durée ou la forme des signaux à utiliser. Une autre approche consiste à limiter l'impact de cette dégradation de l'oxyde tunnel plutôt que la dégradation elle-même. Ces différentes méthodes seront détaillées dans la suite de ce chapitre.

2.2 Etude de l'impact des pulses courts

2.2.1 Bibliographie

De précédentes études ont montré que l'utilisation de signaux très courts (de l'ordre de quelques centaines de ns) par rapport aux signaux habituellement utilisés (de l'ordre de quelques centaines de μs), pouvait apporter une diminution de la dégradation de l'oxyde tunnel lors du cyclage [Irrera'04][Chimenton'06].

En effet, si l'on considère la dynamique de création des pièges à l'intérieur de l'oxyde tunnel lors de l'application d'un champ électrique, il est possible de montrer

qu'il existe une durée seuil avant laquelle les pièges créés ne sont pas stables et peuvent se relaxer de façon spontanée, comme le montre la figure 2.3 [Irrera'04]. En conséquence, si la durée du signal appliqué en programmation est inférieur au temps de génération des défauts stables de l'oxyde, ces pièges relaxeront spontanément et la dégradation sera moins importante. En appliquant une succession de plusieurs signaux courts plutôt qu'un seul signal long, avec la même cible de tension de seuil, nous devrions pouvoir diminuer les niveaux de dégradation observés.

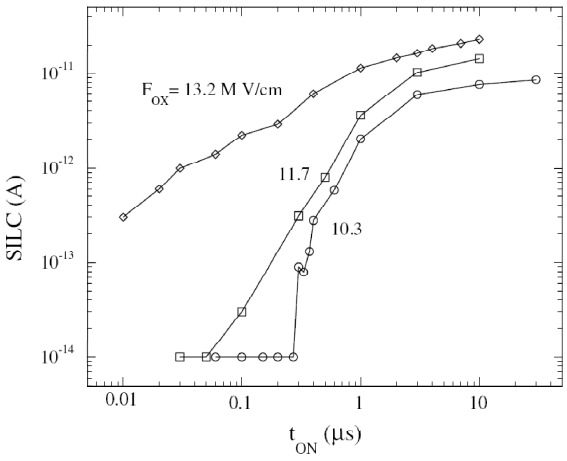


FIG. 2.3 – Valeurs expérimentales du courant de SILC (Stress Induced Leakage Current), fonction de la durée du pulse, selon Irrera [Irrera'04]

Nous avons donc choisi de vérifier l'efficacité de cette méthode de programmation sur différentes cellules mémoires développées par l'entreprise Atmel-Rousset.

2.2.2 Expérience sur la structure "S16"

2.2.2.1 Présentation de la structure "S16"

La structure appelée "S16" dans la suite de ce manuscrit est une mémoire Flash en architecture NAND $90nm$, composée d'une chaîne de 16 cellules mémoires élémentaires, décrite dans la figure 2.4.

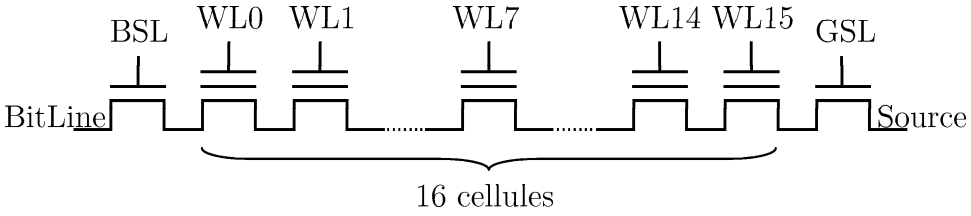


FIG. 2.4 – Structure mémoire S16

2.2.2.2 Faisabilité

Pour limiter au maximum la dégradation de la cellule lors de la phase de programmation, les études précédemment mentionnées préconisent d'utiliser une suite de signaux de la durée la plus faible possible plutôt qu'un seul signal de grande durée. Ici, le temps de plateau des signaux en mode pulsé a pu être diminué jusqu'à $300ns$ pour le signal de programmation avec une bonne maitrise de la forme du signal. Cette valeur est du même ordre ou légèrement inférieure aux temps de création des pièges que l'on peut trouver dans la littérature [Irrera'04]. Les contraintes sont multiples lors de l'application de signaux d'une durée aussi faible sur nos cellules :

-La forme du signal en sortie du générateur doit respecter la forme du signal souhaité, notamment dans la montée qui ne doit être ni trop lente pour garder un signal carré, ni trop rapide pour ne pas engendrer de dépassement provoquant un pic de tension.

-Ce signal doit ensuite pouvoir être acheminé, sans être déformé, jusqu'à la cellule via les câbles, les pointes du banc de mesure et les diverses lignes de métal avant la cellule. Nous pouvons vérifier la bonne géométrie du signal effectivement appliqué sur la cellule en visualisant ce signal à l'oscilloscope. La figure 2.5 présente le signal défini et réellement appliqué au niveau de la pointe permettant le contact sur la structure de test.

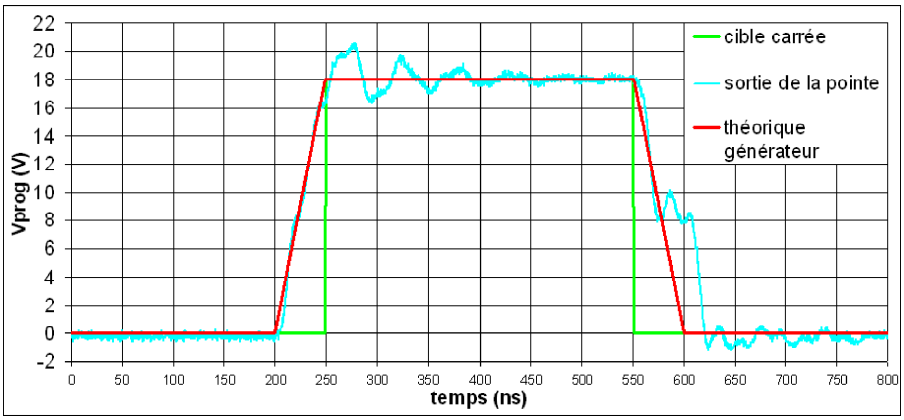


FIG. 2.5 – Visualisation du pulse de $300ns$, carré idéal, théorique pour la définition dans le générateur de signaux arbitraires et mesuré au niveau de la pointe permettant le contact sur la structure de test

Nous avons ainsi pu valider la génération de signaux d'amplitude $18V$ de $300ns$ de plateau avec des temps de montée et de descente de $50ns$.

Nous avons aussi réalisé des mesures de programmation et d'effacement en fonction du temps avec ces signaux très courts, que nous avons ensuite comparées avec des mesures de programmation en fonction du temps en utilisant des signaux plus longs

(de l'ordre de quelques μs ou dizaines de μs) dont nous connaissons de façon certaine la forme au niveau de la cellule. En effet, un signal déformé au niveau de la cellule n'aurait pas la même efficacité de programmation et d'effacement, c'est pourquoi nous avons réalisé ces mesures de programmaiton et d'effacement en fonction du temps.

On obtient alors les courbes de programmation en fonction du temps de la figure 2.6. Avec 40 signaux élémentaires d'une durée de $300ns$ chacun, soit une durée de plateau totale de $12\mu s$, la cellule est passée d'une tension de seuil de $-4V$ à une tension de seuil de $+1,5V$. En utilisant un seul signal de $12\mu s$, nous atteignons également une tension de seuil programmée de $+1,5V$, ce qui montre que nos signaux de $300ns$ s'appliquent correctement au niveau de la cellule.

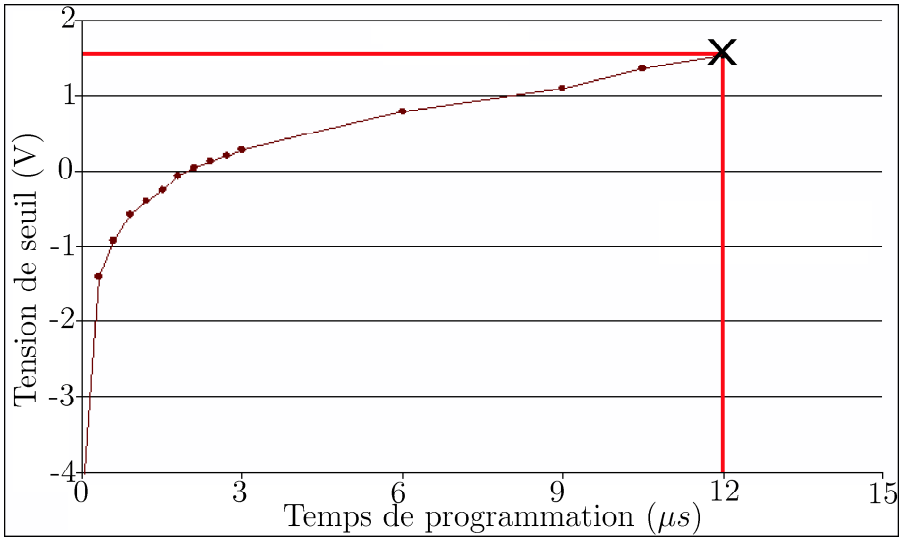


FIG. 2.6 – Programmation effectuée avec des signaux de $300ns$ (•) ou avec un signal de $12\mu s$ (×)

Pour le signal d'effacement, la forme désirée est respectée en sortie de générateur avec un signal de $300ns$ mais la cellule ne s'efface pas correctement comme nous pouvons le voir sur la figure 2.7. Nous pouvons expliquer cela par la difficulté d'établir un signal aussi court sur le substrat de notre cellule, du fait de sa grande surface et par conséquent de sa grande résistance d'accés. Nous avons donc été contraints d'augmenter la durée du signal à $1\mu s$. L'efficacité du signal d'effacement est alors très bonne et bien meilleure que pour 3 pulses de $300ns$ pour une durée totale équivalente. Avec $1\mu s$ de signal d'effacement, la polarisation a le temps de s'établir correctement sur le substrat et l'on retrouve les mêmes cinétiques d'effacement que pour des signaux de plus grande durée. Les signaux d'effacement pour la suite de cette étude utilisant des pulses courts auront donc une durée de $1\mu s$.

Les cyclages, appelés par impulsions, réalisés par la suite utilisent donc :
- en programmation : 40 signaux de $300ns$ à 18V, espacés de $30\mu s$
- en effacement : 3 signaux de $1\mu s$ à 18V, espacés de $30\mu s$

Le repos de $30\mu s$ entre les pulses successifs répond à une contrainte de temps nécessaire au dépiégeage de charges entre deux pulses, définie par les publications citées précédemment au paragraphe 2.2.1 [Irrera'04][Chimenton'06].

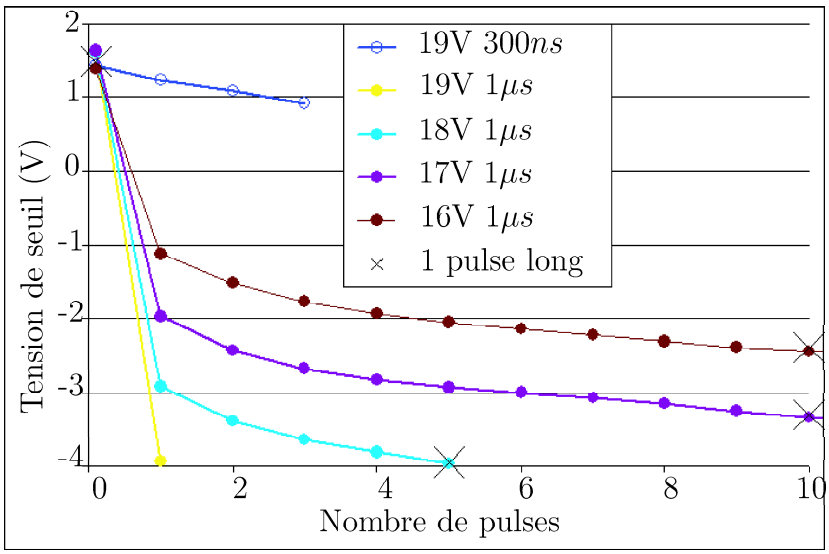


FIG. 2.7 – Effacement effectué avec des signaux de $300ns$ (○), de $1\mu s$ (●) ou avec un seul signal long de même durée totale (×)

Du fait des différences traditionnellement observées selon la position de la cellule dans la chaîne mémoire, nous avons effectué l'étude sur 3 cellules dans la chaîne :
- la première cellule du côté de la Bit Line, correspondant à la Word Line WL0
- une cellule située au milieu de la chaîne, correspondant à la Word Line WL7
- la dernière cellule de la chaîne, située du côté de la Source, correspondant à la Word Line WL15.

2.2.2.3 Résultats

Le mode de cyclage utilisant les signaux courts sera appelé "cyclage par impulsions" tandis que le mode de cyclage utilisant un unique signal de $200\mu s$ à $18V$ en programmation et un unique signal de $1ms$ à $15,5V$ sera appelé "cyclage classique". La figure 2.8 présente les médianes des variations de tensions de seuil sur une dizaine de structures de test, réparties sur l'ensemble de la plaque afin de s'affranchir d'éventuelles variations selon la situation sur la plaque[1], après 100.000 cycles. Ces

[1]On remarque parfois des disparités de mesures entre les structures placées en bord de plaque et celles placées au centre, en raison de variations de grandeurs telles que les épaisseurs d'oxyde

structures de test sont constituées de deux chaînes S16 de Lignes de Bits voisines, situées au milieu d'une matrice d'environ 500×500 chaînes NAND S16.

Nous avons finalement pu établir que dans notre cas l'utilisation de signaux courts n'apporte aucune amélioration au niveau de la dégradation de l'oxyde tunnel, voire même induit un niveau de dégradation plus élevé dans le cas de la cellule WL0. Cela peut s'expliquer par le fait que la durée minimale du signal élémentaire n'a pas pu être abaissée en dessous de $300ns$ pour des raisons de limitation des générateurs de signaux utilisés et que cette durée est très proche de la durée τ_{cr} nécessaire à la création de pièges stables dans l'oxyde tunnel, valant $\tau_{cr} = 300ns$ pour un champ électrique $E_{tun} = 10,8MV/cm$ aux bornes de l'oxyde tunnel [Irrera'04]. Dans notre cas, en utilisant une tension de programmation $V_{prog} = +18V$ aux bornes d'un oxyde tunnel d'une épaisseur $t_{tun} = 8nm$, le champ électrique peut être estimé de façon simple par l'expression :

$$E_{tun} = \frac{V_{tun}}{t_{tun}} = \frac{\alpha_g.V_{prog}}{t_{tun}}$$

où le coefficient de couplage entre la grille de contrôle, sur laquelle est appliquée la tension V_{prog}, et la grille flottante vaut $\alpha_g = 0,52$. Ainsi, le champ électrique aux bornes de notre oxyde tunnel vaut $E_{tun} = 11,7MV/cm$. Le temps de création de pièges stables variant en sens inverse du champ électrique, lorsque nous appliquons des signaux de $300ns$, des pièges stables, voire tous les pièges stables puisque les niveaux de dégradation sont comparables à ceux obtenus avec un unique signal long, ont eu le temps de se créer.

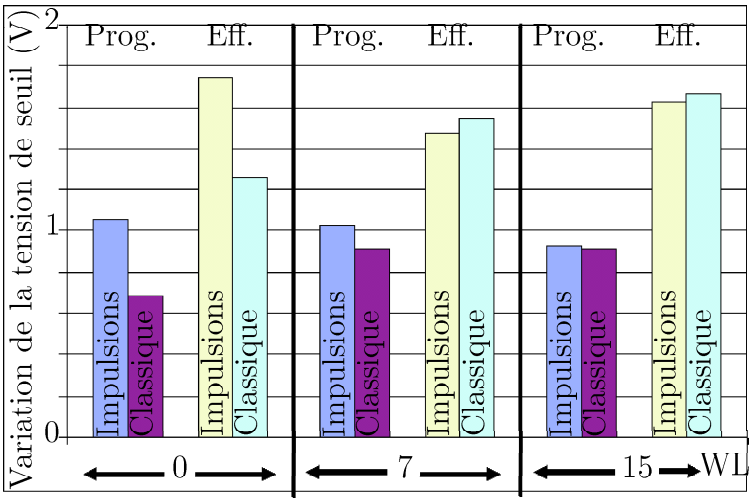


FIG. 2.8 – Comparaison des dégradations des tensions de seuil en fonction de la WL et des signaux utilisés, après 100.000 cycles

2.2.3 Expérience sur la structure "S1"

2.2.3.1 Présentation de la structure "S1"

La structure appelée "S1" dans la suite de ce manuscrit est une mémoire Flash en architecture NAND composée d'une cellule mémoire élémentaire, placée entre deux transistors de sélection et décrite dans la figure 2.9

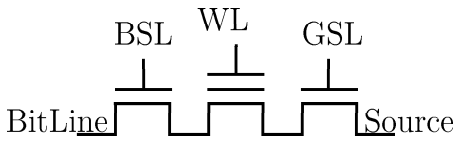


FIG. 2.9 – Cellule mémoire S1

2.2.3.2 Définition des cellules dites "sélectionnées" et "inhibées"

Dans une matrice mémoire en architecture NAND, et quel que soit le nombre de cellules dans la chaîne mémoire, l'ensemble des Words Lines, les grilles des transistors de sélection et la ligne de source sont respectivement reliées entre elles par des lignes de métal. Seules les lignes de Bit sont distinctes et permettent d'adresser une chaîne de cellules mémoires parmi celles contenues dans l'ensemble de la matrice qui peut comporter jusqu'à 2048 chaînes. Dans la matrice mémoire de la figure 2.10, on appelle "cellule sélectionnée" la cellule en cours d'utilisation (en phase de programmation, effacement ou lecture). La cellule dite "inhibée" est une cellule appartenant à une autre chaîne mémoire, donc ayant une autre ligne de bit, mais qui partage sa Word Line avec la cellule sélectionnée. Le terme "inhibée" est dû au fait que lors de la programmation de la cellule sélectionnée par l'application d'une polarisation sur la Word Line, toutes les cellules qui possèdent cette même Word Line devraient théoriquement être programmées. On empêche cette programmation non souhaitée sur ces cellules en les "inhibant" lors de la phase de programmation par une méthode que nous détaillerons dans le chapitre 4.

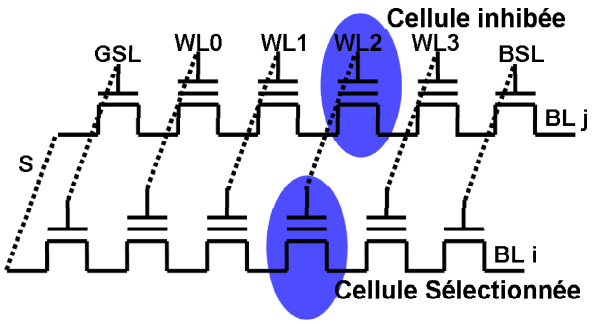


FIG. 2.10 – Matrice mémoire NAND de chaînes à 4 cellules mémoires

Si par exemple dans la matrice mémoire NAND formée de chaînes de 4 cellules mémoires, on considère que la cellule définie par les coordonnées (WL2,BLi) est la cellule "sélectionnée", alors la cellule ayant pour coordonnées (WL2,BLj) est appelée cellule "inhibée".

2.2.3.3 Faisabilité

Les mesures ont dû être réalisées dans le laboratoire de caractérisation électrique de l'entreprise Atmel, ce qui a impliqué un changement de banc de mesure par rapport aux mesures sur la technologie S16, réalisées sur les bancs de mesure de l'IM2NP. La taille minimale des pulses était limitée matériellement par le générateur arbitraire à une durée de plateau de $5\mu s$ du fait de formes de signaux plus complexes, ce qui nécessite un plus grand nombre de points dans le signal pour garantir sa forme finale.

2.2.3.4 Protocole expérimental

La lecture de l'état de la cellule se fait dans le cas présent non pas par une lecture classique de la tension de seuil de la cellule mémoire mais par une lecture du courant à une tension de grille donnée. Au lieu de lire la tension pour laquelle on atteint un courant donné, on lit le niveau de courant atteint à une tension donnée, ce qui est tout à fait équivalent mais correspond à une contrainte imposée par le produit dans lequel l'état de la cellule est détecté par un comparateur de courant.

Ainsi, les niveaux initiaux lors du cyclage sont des niveaux de courant que l'on fixe pour toutes les structures à mesurer et ce quelles que soient l'amplitude et la durée des signaux utilisés. Les structures de test utilisées dans cette partie sont similaires à celles utilisées dans le paragraphe 2.2.2.3 hormis que les chaînes ne comportent qu'une seule cellule au lieu de seize. Lors de cette étude nous cherchons à mettre en évidence une éventuelle diminution de la dégradation induite par 100.000 cycles en jouant sur la durée du signal de programmation élémentaire, mais pour une charge totale constante.

Le tableau 2.1 définit le nombre de pulses utilisés lors de la phase de cyclage. Ce nombre de pulses est déterminé initialement pour chaque cellule par une mesure préalable de programmation et d'effacement qui détermine le nombre exact de signaux à appliquer pour garantir une fenêtre (en courant) initiale constante.

Durée d'un pulse (μs)	**Nombre de pulses**
200	1
100	2-3
40	4-6
5	40-50

Tab. 2.1 – Nombre de pulses de programmation sur cellule S1

Avant d'analyser les résultats de dégradation en fonction de la durée des pulses utilisés, nous pouvons déjà remarquer dans le tableau 2.1 que l' "asservissement" du nombre de pulses sur une valeur-cible de courant (mais le problème serait le même pour une valeur-cible en tension) pose des problèmes lorsque la durée du pulse élémentaire est élevée. En effet, chaque pulse appliqué implique un ΔV_T qui augmente avec la durée du pulse élémentaire. Deux remarques apparaissent à partir de ce constat :

- une dispersion de la tension de seuil autour de la valeur-cible d'autant plus grande que la durée du pulse élémentaire est grande,
- un risque important de sur-programmation lorsque la durée du pulse élémentaire augmente

En revanche plus le pulse élémentaire est fin, plus la distribution est fine, au détriment d'un nombre de pulses et donc d'une durée totale de programmation plus dispersés.

2.2.3.5 Résultats

La figure 2.11 représente la perte de tension de seuil après une application de 100.000 cycles, suivant la durée du pulse élémentaire. Il semble d'après les valeurs reportées que dans le cas de notre cellule "S1", l'utilisation de signaux courts de l'ordre de $5\mu s$, induit des niveaux de dégradation légèrement plus faibles, de l'ordre de $100mV$, sur la cellule sélectionnée par rapport aux niveaux de dégradation mesurés avec des signaux longs. En revanche en relevant au cours de ces cyclages la dégradation de la cellule inhibée, telle que définie au paragraphe 2.2.3.2, nous attendent des niveaux de dégradation plus élevés pour de faibles temps de pulses comme le montre la figure 2.12. Si les faibles temps de pulses améliorent d'environ $100mV$ les niveaux de dégradation de la cellule sélectionnée, la contrepartie à payer au niveau des cellules inhibées semble aller à l'encontre de l'utilisation de tels signaux pour nos cellules en raison de la hausse de $100mV$ et $250mV$ des valeurs de décalage de V_T programmé et effacé respectivement. Une étude détaillée pour l'explication de la dégradation des cellules inhibées sera menée dans le chapitre 4.

2.2.4 Discussion

Le fait que les niveaux de dégradation observés, aussi bien sur la cellule S16 que sur la cellule S1, ne soient pas plus faibles lors de l'utilisation de signaux de courte durée peut s'expliquer par le fait que nos signaux avaient une durée supérieure au temps de création des pièges stables et par conséquent les niveaux de dégradation sont comparables. Il ne nous est cependant pas possible, à la fois matériellement avec nos bancs de mesures mais également par la suite dans le produit final, d'obtenir des durées de pulse inférieures à celles utilisées dans l'étude précédente. Cette méthode de programmation à l'aide de signaux très courts n'a donc pas été retenue par la suite lors du développement de ces technologies mémoires.

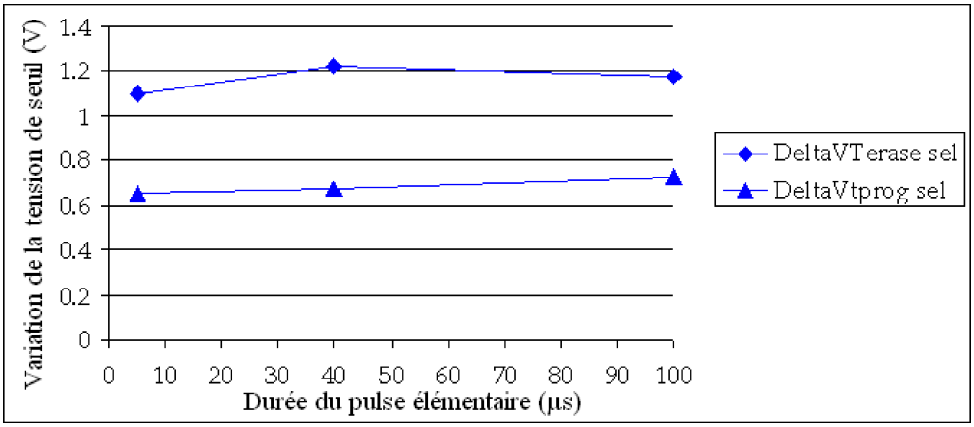


FIG. 2.11 – Variation de la dégradation de la tension de seuil de la cellule sélectionnée avec la durée du pulse élémentaire

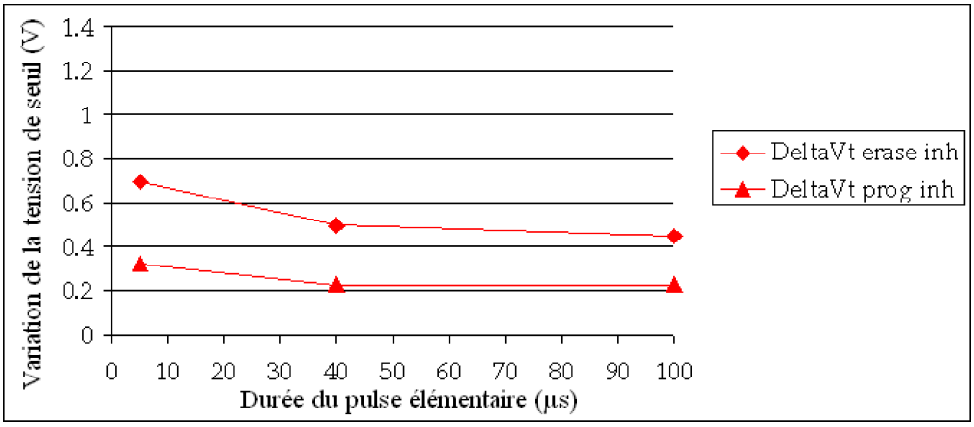


FIG. 2.12 – Variation de la dégradation de la tension de seuil de la cellule inhibée avec la durée du pulse élémentaire

Le problème de la dégradation des cellules inhibées qui augmente avec la diminution de la durée du signal élémentaire de programmation sera à nouveau abordé au chapitre 4.

2.3 Théorie des signaux optimisés

En vue d'améliorer les performances en endurance des cellules, une autre approche serait de minimiser la valeur maximale du champ électrique appliqué. Il est possible de définir une nouvelle forme de signaux appelés "optimisés" qui dégradent moins

la cellule au cours du cyclage [Canet'01]. On définit le signal optimisé de la façon suivante :

Soit le schéma électrique 2.13, la source de courant I_{FN} sera non-nulle uniquement lors de l'injection des charges dans la grille flottante. En appliquant la loi de Kirchoff au nœud de la grille flottante, on obtient la relation :

$$I_{FN} = (C_d + C_{ox} + C_{pp}) \times \frac{dV_{fg}}{dt} - (\frac{dV_{cg}}{dt} \times C_{pp}) \tag{2.7}$$

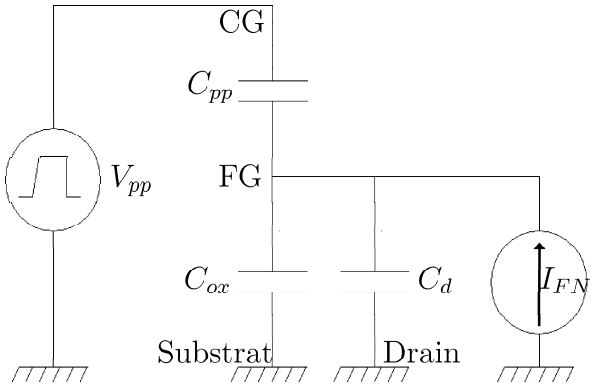


FIG. 2.13 – Schéma électrique équivalent en phase de programmation

Hors injection, I_{FN} est nul et la résolution de l'équation (2.7) conduit à la définition du coefficient de couplage capacitif :

$$K_e = \frac{\frac{dV_{fg}}{dt}}{\frac{dV_{cg}}{dt}} = \frac{C_{pp}}{C_d + C_{ox} + C_{pp}} \tag{2.8}$$

Lors de l'injection, nous pouvons considérer que celle-ci a lieu pour des tensions comprises entre $8V$ et $9V$ dans le cas d'une épaisseur d'oxyde de l'ordre de $8nm$. Dans cette région de la courbe, nous pouvons utiliser l'approximation linéaire du courant, de pente g :

$$I_{FN} = g \times (V_s - V_{fg}) \tag{2.9}$$

En résolvant alors l'équation (2.7), on obtient :

$$V_{fg} = V_s + \frac{r \times C_{pp}}{g} \times [1 - \exp(\frac{-g \times (t - t_s)}{C_d + C_{ox} + C_{pp}})] \tag{2.10}$$

où $r = \frac{dV_{cg}}{dt}$.

L'expression de la valeur maximale de la tension de grille flottante s'écrit alors :

$$V_{fgmax} = V_s + \frac{r \times C_{pp}}{g} \tag{2.11}$$

On peut alors remarquer que si l'on minimise cette valeur on diminue également la valeur maximale du champ électrique aux bornes de l'oxyde tunnel. Cela équivaut à réduire V_s, soit le courant d'injection. On fixe ainsi la pente g. On peut également minimiser la valeur de la rampe r appliquée sur la grille de contrôle.

La figure 2.14 définit le signal optimisé qui est constitué par :

- une rampe abrupte jusqu'à une tension V_{Low} inférieure à la tension de seuil d'injection
- une rampe la plus faible possible jusqu'à la tension V_{High} qui donne la même quantité de charges injectées et par conséquent la même tension de seuil.

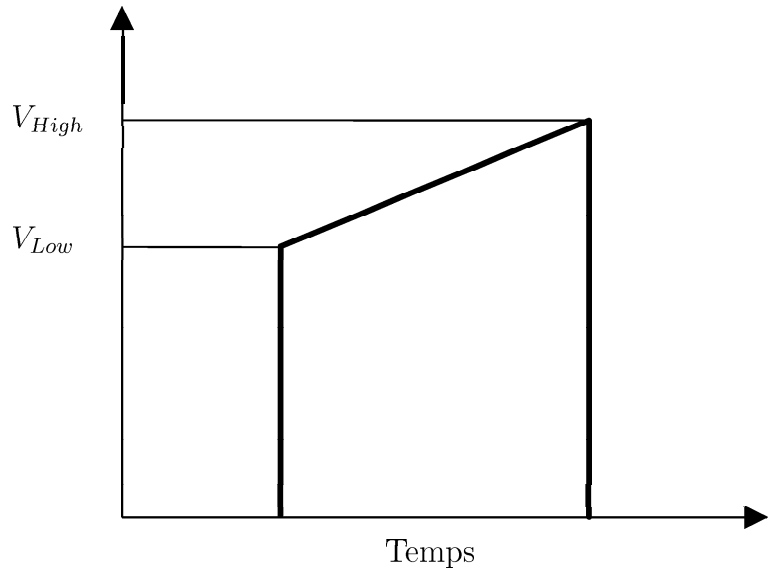


FIG. 2.14 – Allure du signal optimisé

Des résultats de simulation Eldo, donnés en Annexe **I**, montrent que l'utilisation du signal optimisé correspond à une injection à courant constant et que le champ électrique maximum diminue de près de $1MV/cm$ [Canet'01]. La quantité de charges injectées étant égale à l'intégrale du courant, la charge est injectée de façon continue.

Cette voie d'amélioration a été explorée lors d'études antérieures à ce travail de thèse. Nous présentons des résultats obtenus lors de ces précédentes études. Nous n'avons pas directement appliqué cette méthode aux cellules mémoires considérées dans ce manuscrit mais elle faisait partie d'une des possibilités que nous nous réservions d'utiliser au cours de notre travail.

2.4 Algorithme de programmation "intelligent"

2.4.1 Principe de la programmation intelligente

Au lieu de définir un premier signal pour la programmation et un second pour l'effacement de nos cellules, valables quels que soient la cellule et son état de dégradation, il est possible d'utiliser des signaux qui s'adaptent à ces variations et qui garantissent l'état final de la cellule après la phase de programmation ou d'effacement. Ainsi un algorithme de programmation dit "intelligent" peut être développé à partir d'une succession de pulses relativement courts, entrecoupés de lectures permettant de vérifier l'état de la cellule. Ainsi, si la cellule n'est pas encore considérée comme étant programmée, on continue la programmation, sinon on arrête. La figure 2.15 présente cet algorithme sous forme d'un schéma fonctionnel.

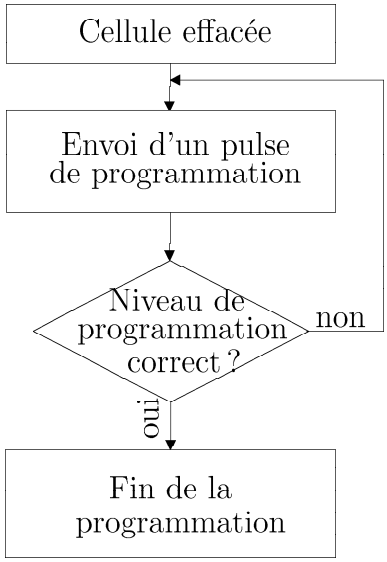


FIG. 2.15 – Principe de l'algorithme de programmation dit "intelligent"

2.4.2 Mise en œuvre de la programmation intelligente

Nous remarquons sur les figures 2.16 à 2.19, qui présentent des étapes d'une programmation intelligente avec des signaux de $5\mu s$ et d'amplitude $17V$, l'intérêt de cette méthode de programmation de la cellule qui garantit la tension de seuil des niveaux "programmé" et "effacé", quel que soit le niveau de dégradation. En revanche, cela complexifie la circuiterie de gestion des signaux qui effectuent la vérification du niveau de la cellule entre les phases de programmation. Ceci a également un "coût" d'un point de vue temps de programmation car entre deux signaux appliqués sur la cellule, il faut faire une mesure de son état. Le temps total de programmation est donc beaucoup plus grand que la somme des signaux effectivement appliqués sur la

cellule. Il faut donc mettre en place une stratégie efficace afin de savoir quand une vérification intermédiaire est nécessaire ou quand elle est superflue.

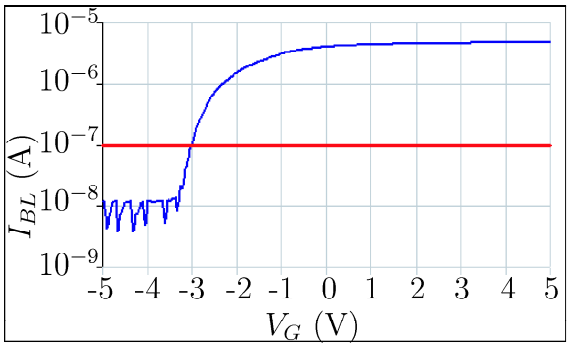


FIG. 2.16 – Programmation intelligente d'une cellule - avant la phase de programmation

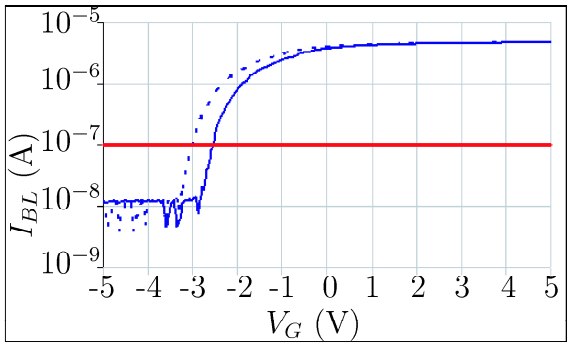


FIG. 2.17 – Programmation intelligente d'une cellule - après 1 pulse de programmation

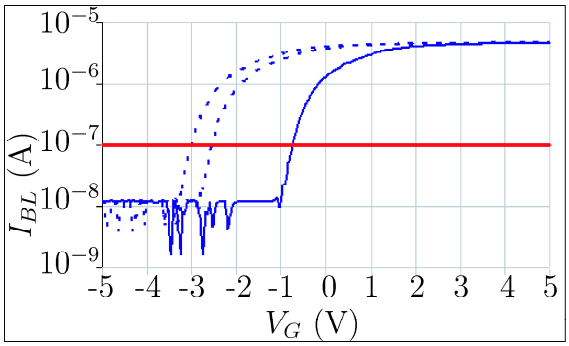


FIG. 2.18 – Programmation intelligente d'une cellule - après 10 pulses de programmation

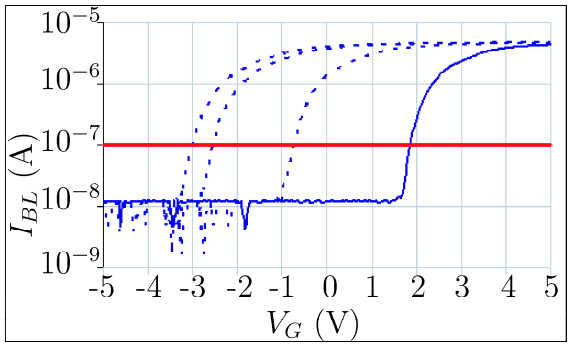


FIG. 2.19 – Programmation intelligente d'une cellule - après 40 pulses de programmation

2.4.3 Cyclages en programmation intelligente

Nous avons choisi d'expérimenter cette méthode de programmation intelligente sur les structures de test S16. La figure 2.20 présente l'évolution des tensions de seuil programmée et effacée en fonction du nombre de cycles, avec des signaux élémentaires de $40\mu s$ et à une tension de $17V$. Cette méthode de programmation "intelligente" garantit certes le niveau de courant de la cellule et donc la tension de seuil au cours du cyclage mais au détriment de la durée d'effacement qui augmente avec le nombre de cycles réalisés. L'efficacité diminuant lorsque le nombre de cycles augmente, il faut appliquer un plus grand nombre de signaux pour atteindre le même état de la cellule. Le cyclage ne se traduit donc plus par un rapprochement des niveaux "0" et "1" avec

un risque de mélange des deux états comme le montre la figure 2.20, mais par une augmentation du temps d'effacement, visible sur la figure 2.21, qui peut à son tour devenir un point critique et compromettre le bon fonctionnement de la cellule.

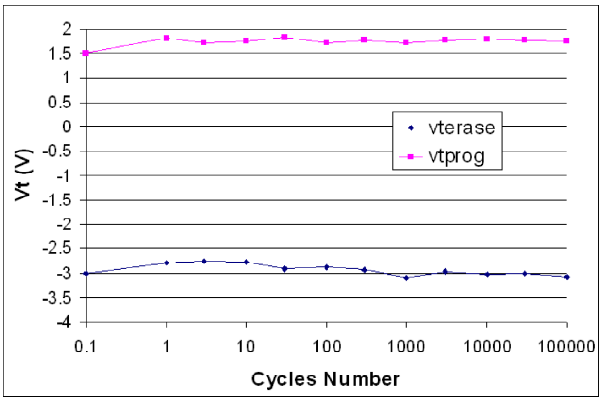


FIG. 2.20 – Evolution des tensions de seuil écrite et effacée avec le nombre de cycles en programmation intelligente

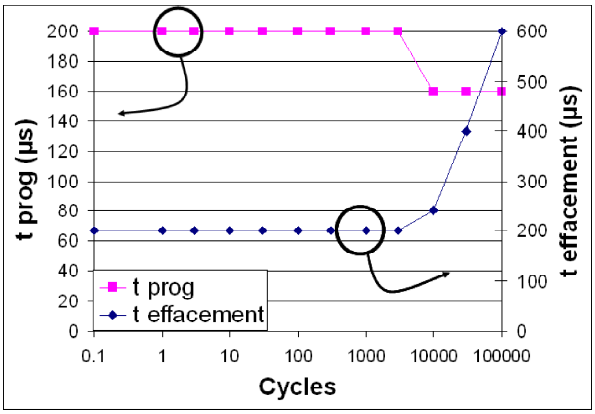


FIG. 2.21 – Evolution des durées totales de programmation et d'effacement avec le nombre de cycles en programmation intelligente

Cette méthode peut donc présenter des intérêts dans certaines applications où le point critique est la bonne discrimination des états logiques, plutôt que les temps de programmation et d'effacement. En revanche, nous n'utiliserons pas dans la suite de ce manuscrit cet algorithme car il nous prive d'un des moyens les plus simples et les plus courants d'observer la dégradation des oxydes, la mesure des décalages des tensions de seuil. De plus, cette méthode garantit la fenêtre de programmation en augmentant le nombre de pulses d'effacement. La cellule subit donc davantage de

signaux, ce qui induit davantage de dégradations qui se traduisent généralement par des pertes en rétention après cyclage plus importantes.

2.5 Conclusion

Lors de ce deuxième chapitre, nous avons abordé différentes méthodes de programmation ayant pour objectif de réduire la dégradation des cellules mémoires lors des phases répétées de programmation/effacement. Nous avons dans un premier temps présenté une étude sur l'utilisation de pulses très courts dont nous n'avons pas pu montrer qu'elle apportait une réelle amélioration sur deux structures Flash en architecture NAND. Nous avons expliqué ce résultat par la limitation de la durée minimale du signal élémentaire qui, pour des raisons matérielles, n'a pas pu être diminuée en-dessous de la durée caractéristique de formation de pièges stables dans l'oxyde tunnel. Nous proposons ensuite le principe de l'optimisation de la forme des signaux utilisés, visant à diminuer le champ électrique maximal aux bornes de l'oxyde tunnel et à injecter les charges dans la grille flottante de façon continue. Une dernière méthode de programmation consiste en un algorithme de programmation dans lequel le signal de programmation évolue avec la dégradation de la cellule pour garantir une fenêtre de tensions de seuil constante au cours du cyclage. Néanmoins, l'augmentation de la durée d'effacement peut compromettre le bon fonctionnement du produit.

Références bibliographiques du chapitre 2

[Jepson'77] K.O. Jepson, C.M. Svensson
"Negative bias stress of MOS devices a high electric fields and degradation of NMOS devices"
Journal of Applied Physics, vol. 48, pp.2004-2014, 2001.

[Euzent'81] B. Euzent, N. Boruta, J. Lee, C. Jenq
"Reliability aspects of a floating gate EEPROM"
Proceedings of IRPS, pp.11-16, 1981.

[Canet'01] P. Canet, R. Bouchakour, N. Harabech, P. Boivin, J.M. Mirabel, C. Plossu
"Improvement of EEPROM cell reliability by optimization of signal programming"
Journal of Non-Crystalline Solids, 280, pp.116-121, 2001.

[Baboux'03] N. Baboux, C. Busseret, C. Plossu, S. Burignat, B. Ballant, P. Boivin
"Towards a model linking tunnel oxide degradation to programming window closure in EEPROM cells"
Journal of Non-Crystalline Solids, vol. 322, pp.240, 2003.

[Irrera'04] F. Irrera, T. Fristachi, D. Caputo, B. Ricco
"Optimising flash memory tunnel programming"
Microeletronic Engineering, 72, pp.405-410, 2004.

[Lee'04] J.D. Lee, J.H. Choi, D. Park, K. Kim
"Effects of interface trap generation and annihilation on the data retention characteristics of flash memory cells"
IEEE Transactions on Device and Materials Reliability, vol.4, no.1, 2004.

[Denais'05] M. Denais, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, D. Roy, A. Bravaix
"New perspectives on NBTI in Advanced technologies : modeling and characterization"
European Solid-State Device Research Conference, 2005.

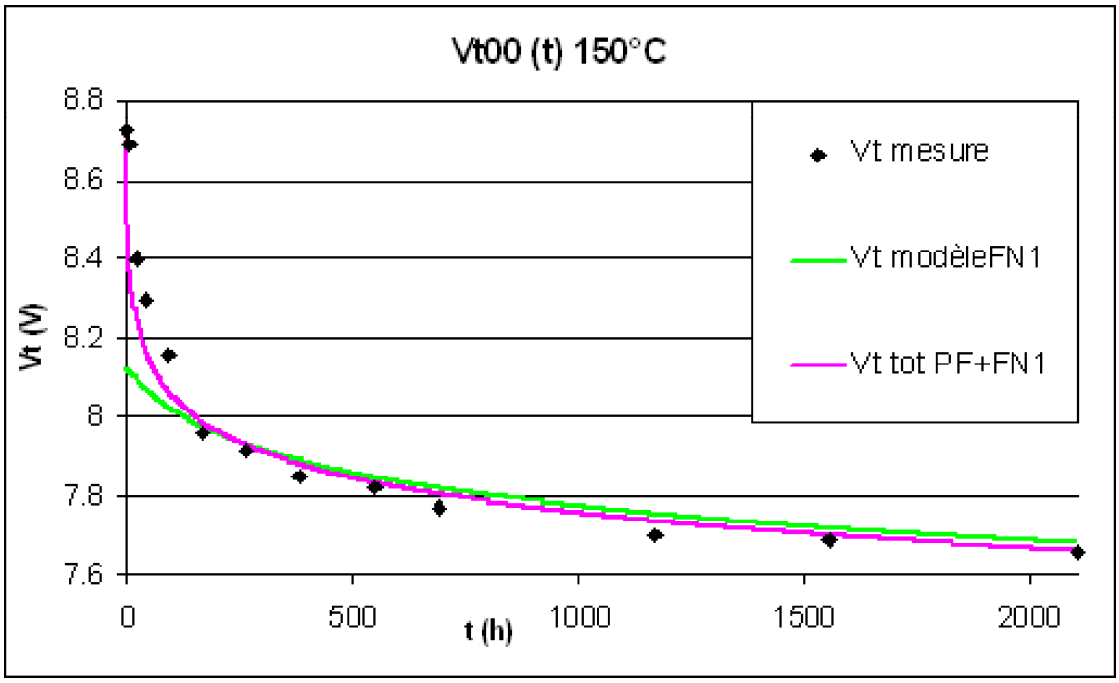
[Chimenton'06] A. Chimenton, F. Irrera, P. Olivo
"Ultra-short pulses improving performance and reliability in Flash memories"
Non-Volatile Semiconductor Memory Workshop Technical Digest, pp.46-47, 2006.

[Tseng'06] J.M.Z. Tseng, T. Pédron
"A new method to extract gate coupling ratio and oxide trapped charge in flash memory cell"
Microelectronic Engineering, 83, pp.218-220, 2006.

[Bénard'08] C. Bénard
"Etude des phénomènes de dégradation des transistors MOS de type porteurs chauds et negative bias temperature instability (NBTI)"
Thèse de doctorat, Université de Provence, 2008.

Chapitre 3

Fiabilité des Mémoires Flash



Sommaire

Le niveau d'intégration croissant des dispositifs mémoires non volatiles de type *Flash* induit une diminution des charges stockées dans la grille flottante et par voie de conséquence l'obligation de maintenir des courants de fuite à un niveau extrêmement faible pour assurer la non volatilité de ces mémoires et garantir la bonne rétention des données. Typiquement, une cellule Flash classique sur le marché actuellement doit pouvoir maintenir les données pendant une dizaine d'années à température ambiante, ce qui équivaut à une perte d'environ un électron par jour. En effet, si l'on considère une charge stockée dans la grille flottante $Q_{fg} \approx 1 \times 10^{-15}C$, cela correspond à 6200 électrons de charge élémentaire $q = 1,6 \times 10^{-19}C$. Pour garder au moins la moitié de ces électrons stockés après une période de 10 ans, soient 3600 jours, cela revient bien à perdre moins d'un électron par jour. De plus, la nécessité de réduire la tension d'alimentation et la consommation doit conduire à la diminution des épaisseurs d'oxyde, ce qui augmente la probabilité de fuite des cellules. La résistance au cyclage est aussi impactée par cette évolution. Les opérations répétées d'effacement et d'écriture, le cyclage, peuvent entraîner, sur certaines cellules du bloc mémoire, l'apparition d'un courant de fuite suffisant pour compromettre le fonctionnement du bloc entier. Ce sont davantage les problèmes de fiabilité des mémoires qui constituent un frein au processus de réduction de la taille et de la consommation des circuits, que les problèmes technologiques directement liés à la réduction des dimensions [SEMATECH'03][ITRS'05]. En vue d'améliorer les performances des mémoires, il sera nécessaire d'obtenir une meilleure compréhension des mécanismes limitants, pour prévoir et anticiper au plus tôt les problèmes de fiabilité. Les mécanismes de fuite seront principalement étudiés du fait de leur importance primordiale dans la rétention des données.

Le but de ce chapitre est de donner une vision d'ensemble des mécanismes physiques de conduction et de dégradation intervenant dans cet aspect de la fiabilité des mémoires *Flash* par l'interprétation des caractérisations électriques de la technologie étudiée et par un modèle de fuite de charges correspondant aux mesures réalisées. Avant cela nous présenterons le principe des cellules multi-niveaux, cadre de cette étude, ainsi que l'ensemble des conditions électriques utilisées. Suite à l'exposé des mesures réalisées, nous proposerons une modélisation des mécanismes de fuites mettant en évidence leur nombre et leur nature. Nous proposerons également une modélisation correspondant à un "gain" apparent de charges qui survient parfois lors des mesures en rétention.

3.1 Protocole Expérimental

3.1.1 NOR Multi-Level

La réduction des coûts de fabrication étant un élément-clé dans le domaine des semiconducteurs, l'idée de pouvoir stocker dans une seule cellule standard plusieurs bits au lieu d'un seul présente un intérêt évident [Modelli'01]. Pour définir plusieurs niveaux de tensions de seuil sur une seule cellule, soit la marge entre les niveaux doit être réduite, soit la fenêtre de programmation totale doit être élargie. Dans le premier cas, des problèmes de discrimination de niveaux peuvent apparaître tandis que dans le second cas, le champ électrique dans les diélectriques est augmenté avec des impacts négatifs sur les caractéristiques en cyclage et en rétention [SIMATECH'03].

L'entreprise ATMEL a ainsi développé sur une technologie Flash 130nm en architecture NOR[1] des cellules multibits, permettant de stocker deux bits sur une seule cellule (soient quatre niveaux de V_T : (00), (01), (10) et (11) au lieu des deux habituels, (0) et (1)). Ces quatre niveaux sont équirépartis en courant, c'est-à-dire que l'écart de courant obtenu entre chacun de ces niveaux est le même. Ceci donne des niveaux de V_T, regroupés dans le tableau 3.1.

Niveau	**Valeur de V_T**
(00)	$\approx 8,5V$
(01)	$\approx 6,5V$
(10)	$\approx 5,5V$
(11)	$\approx 2,5V$

TAB. 3.1 – Les quatre niveaux de V_T

Le fait de placer quatre niveaux au lieu de deux dans la même structure rend la discrimination de ces niveaux plus sensible à la perte de charges, le choix ayant été fait de ne pas agrandir la fenêtre de programmation. Cette étude doit donc permettre de vérifier la tenue en rétention des données sur quatre niveaux dans une seule cellule.

Ces quatre niveaux sont fixés par la technologie qui sera utilisée tout au long de ce chapitre.

3.1.2 Description des signaux utilisés

Tous les wafers[2] ont été cyclés à 100.000 cycles avec les mêmes conditions, à la fois de polarisation et de durée, entre les niveaux (00) et (11) qui sont les niveaux

[1] Nous rappelons qu'une description de la cellule mémoire Flash, ainsi que des mécanismes de programmation est disponible au paragraphe 1.1.3.5

[2] Au cours de l'étude 4 wafers seront utilisés en vue de mesures en rétention à 4 températures différentes, 25°C, 85°C, 125°C et 150°C.

les plus éloignés correspondant respectivement aux tensions de seuil $8,5V$et $2,5V$, ce qui équivaut à la condition la plus dégradante car les champs mis en jeu sont les plus élevés lors des phases de programmation et d'effacement.

3.1.2.1 Conditions de Programmation et d'Effacement

Les conditions de polarisation utilisées pour obtenir ces niveaux extrêmes lors de l'étude sont données dans les tableaux 3.2 et 3.3. Les signaux appliqués sont des signaux carrés d'une durée de $4\mu s$ en programmation et de $10ms$ en effacement, dont les chronogrammes sont présentés sur les figures 3.1 et 3.2. Nous pouvons remarquer que lors de la programmation, le substrat est polarisé négativement à $-2V$, ce qui est caractéristique de l'utilisation d'un mécanisme appelé "CHISEL" qui a été présenté dans la figure 1.21 page 34 et qui renforce l'efficacité de la programmation par injection secondaire d'électrons chauds. L'effacement se fait quant à lui par l'utilisation du courant Fowler-Nordheim.

Contacts	Tensions ($4\mu s$)
Drain	3,75V
Grille de Contrôle	8,4V
Source	0V
Substrat	-2V

Tab. 3.2 – Conditions de Programmation

Contacts	Tensions ($10ms$)
Drain	8V
Grille de Contrôle	-8V
Source	8V
Substrat	8V

Tab. 3.3 – Conditions d'Effacement

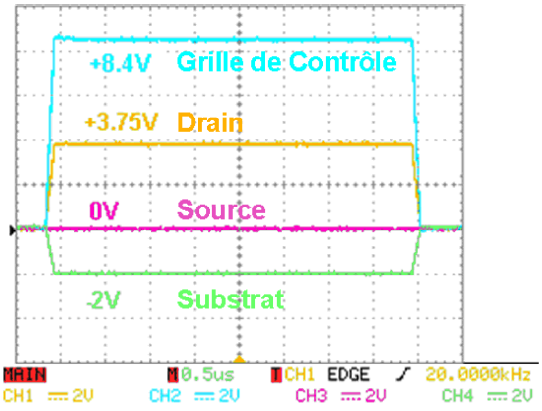


Fig. 3.1 – Chronogramme des signaux de programmation

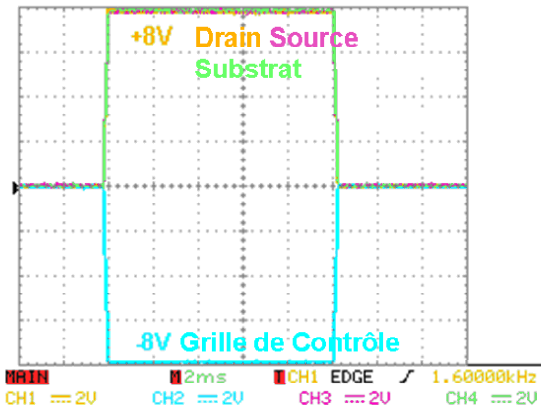


Fig. 3.2 – Chronogramme des signaux d'effacement

3.1.3 Présentation des résultats

Au total quatre wafers ont ainsi été cyclés, à raison d'environ 30 sites par wafer. Chaque site correspond à une structure de test avec une cellule mémoire isolée.

Les différents wafers seront par la suite appelés **W1**, **W2**, **W3** et **W4**.

Nous pouvons donc tracer sur les figures 3.3 à 3.6 le comportement moyen des cellules pour chaque wafer cyclé.

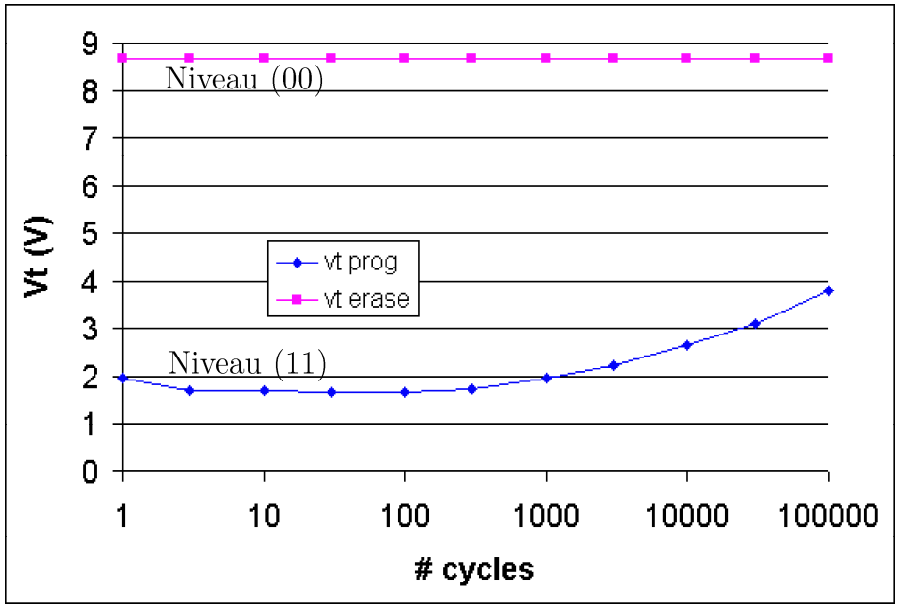


FIG. 3.3 – Cyclage du Wafer **W1**

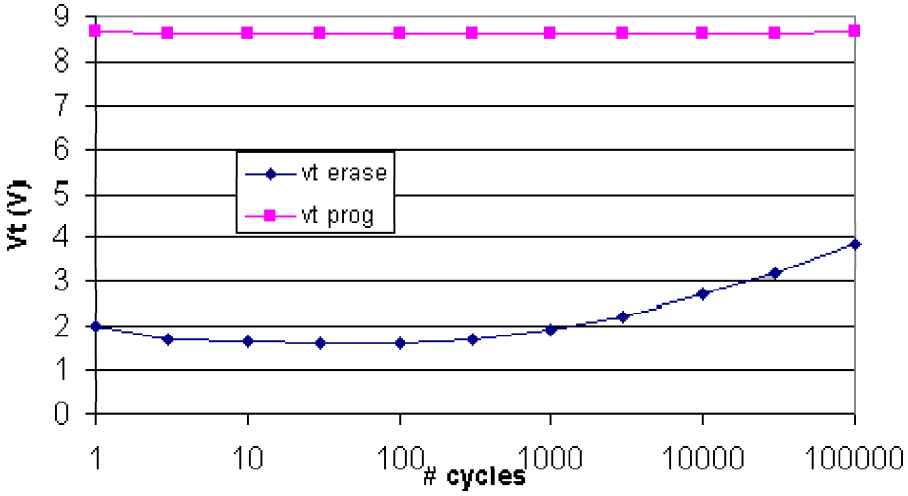


FIG. 3.4 – Cyclage du Wafer **W2**

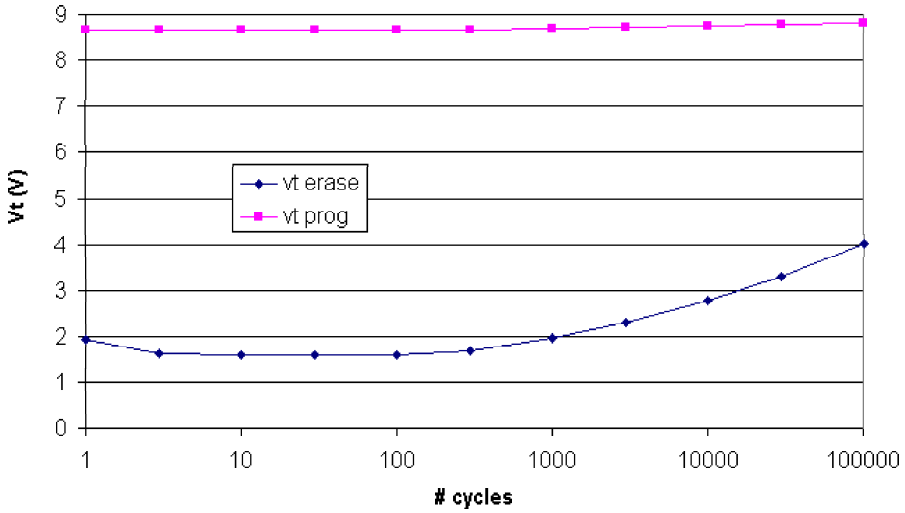


FIG. 3.5 – Cyclage du Wafer **W3**

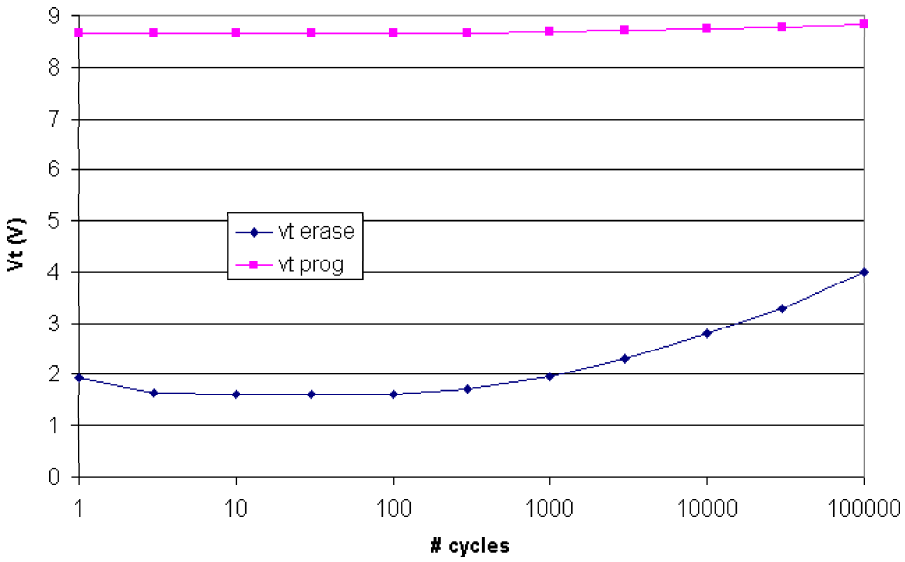


FIG. 3.6 – Cyclage du Wafer **W4**

Le tableau 3.4 résume l'ensemble des différents écarts des V_T programmés, des V_T effacés et des fenêtres de programmation moyennes sur chacun des quatre wafers précédemment cyclés définis comme suit :

$$\begin{aligned} \Delta V_{T_{erase}} &= V_{T_{erase_{final}}} - V_{T_{erase_{initial}}} \\ \Delta V_{T_{prog}} &= V_{T_{prog_{final}}} - V_{T_{prog_{initial}}} \\ \Delta V_{T_{fenetre}} &= \Delta V_{T_{erase}} - \Delta V_{T_{prog}} \end{aligned}$$

Wafers	$\Delta V_{T_{erase}}$	$\Delta V_{T_{prog}}$	$\Delta V_{T_{fenetre}}$
1	1.84V	0.00V	1.84V
2	1.84V	-0.03V	1.87V
3	2.06V	0.16V	1.90V
4	2.05V	0.20V	1.85V

TAB. 3.4 – Comparaison des résultats de cyclage des différents wafers

Nous observons un $\Delta V_{T_{erase}}$ élevé, proche de $+2V$, ainsi qu'un $\Delta V_{T_{prog}}$ positif, ce qui est caractéristique de la présence d'états d'interface, dont le phénomène est décrit au paragraphe 2.1.2. Nous pouvons remarquer que pour tous les wafers, les niveaux de dégradation liés au cyclage sont équivalents ce qui évite d'introduire un biais dans la suite de l'étude. Tout écart éventuel observé dans la suite de cette étude sera donc imputable aux conditions ultérieures de mesures et pourra être décorrélé du cyclage initial.

3.2 Rétention

Dans une cellule programmée, la grille flottante chargée négativement induit un faible champ électrique (de 3 à $4MV/cm$) à travers l'oxyde de grille, dirigé vers la grille flottante, sous l'effet duquel les électrons qui sont emmagasinés dans la grille flottante ont tendance à fuir. La perte de charges due à ce champ résiduel a longtemps été considérée, à tort, comme étant négligeable. A titre d'exemple, sans prendre en compte cette fuite, la durée de vie d'une cellule non défectueuse a été estimée à un million d'années pour une température de fonctionnement de $70°C$, ce qui représente une perte de beaucoup moins d'un électron par jour et ne reflète pas la réalité [Shiner'80][Mielke'83]. Afin de garantir la bonne discrimination de nos quatre niveaux stockés après 10 ans à température ambiante, nous avons dû recourir à un test en rétention à des températures plus élevées afin d'accélérer le phénomène de perte de charges et d'en identifier le mécanisme. Les divers mécanismes présents dans la littérature seront présentés avant de poursuivre notre étude.

3.2.1 Mécanismes de fuite à travers un oxyde

3.2.1.1 Les chemins possibles de fuite de charges

Les électrons stockés dans la grille flottante peuvent fuir par trois chemins principaux :

- à travers l'oxyde interpoly ;
- à travers l'oxyde de grille ;
- à travers l'isolation latérale.

3.2.1.2 A travers l'isolation latérale

Des études antérieures ont montré que cette fuite pouvait être considérée comme négligeable [Mazoyer'92][Candelier'97]. En effet, l'épaisseur importante de ces isolations latérales, supérieure à $90nm$, conjuguée à une nitruration de ces oxydes latéraux, conduit à des transparences tunnel très faibles où un électron n'a quasiment aucune probabilité de traverser l'oxyde.

3.2.1.3 A travers l'oxyde interpoly

Dans le cas d'un oxyde interpoly de type ONO[3], ce qui est le cas des technologies que nous avons employées, la perte de charges causée par cette couche est principalement due à la moins bonne qualité de l'ONO par rapport à l'oxyde de champ car l'ONO est déposé sur du polysilicium, la grille flottante, mais aussi à la présence d'angles aigus dans la géométrie de la structure, ce qui augmente localement la valeur du champ électrique.

3.2.1.4 A travers l'oxyde de grille

Pour des températures élevées, la perte de charge à travers l'oxyde de grille a été expliquée à l'aide de mécanismes tels que l'émission thermoionique [Nozawa'82] ou l'effet tunnel Fowler-Nordheim. Toutefois, Nozawa a trouvé une hauteur de barrière irréaliste ($\approx 1,5eV$) en utilisant ce modèle sur des caractéristiques expérimentales de perte de charges. Par ailleurs, on a longtemps admis que la perte de charges par effet tunnel de type Fowler-Nordheim à travers l'oxyde de grille était négligeable, le courant tunnel correspondant à une barrière de potentiel de $3eV$ étant très faible à bas champ [Crisenza'91]. L'importance de cette fuite tunnel a été réévaluée, en particulier à haute température [Papadas'95], suite aux travaux montrant la forte dépendance du courant Fowler-Nordheim avec la température [Pananakakis'95].

3.2.1.5 Mécanismes intrinsèques

Nous appelons mécanismes intrinsèques, les mécanismes inévitables car directement liés aux caractéristiques du diélectrique parfait. Les deux mécanismes intrinsèques de pertes de charges pouvant intervenir dans les mémoires *Flash* sont le courant tunnel direct et le courant Fowler-Nordheim [FowlerNordheim'28], déjà décrits dans le chapitre 1. Ces deux mécanismes diffèrent par la forme du diagramme de bandes, trapézoïdale dans le premier cas et triangulaire dans le second (cf. figure 1.22 page 35). Les expressions des densités de courant sont les suivantes :

$$J = AE_{ox}^2 \exp\{-\frac{K}{E_{ox}}\left[(\frac{\phi_0}{q})^{3/2} - (\frac{\phi_0}{q} - E_{ox}.t_{ox})^{3/2}\right]\} \tag{3.1}$$

lorsque $q.V_{ox} < \phi_0$

[3]Oxyde-Nitrure-Oxyde

$$J = AE_{ox}^2 \exp\left[-\frac{K}{E_{ox}}(\frac{\phi_0}{q})^{3/2}\right] = AE_{ox}^2 \exp\left[-\frac{B}{E_{ox}}\right] \quad (3.2)$$

lorsque $q.V_{ox} \geq \phi_0$

avec V_{ox} la différence de potentiels aux bornes de l'oxyde de grille, t_{ox} l'épaisseur de l'oxyde de grille, $E_{ox} = \frac{V_{ox}}{t_{ox}}$ le champ dans cet oxyde de grille, ϕ_0 la hauteur de barrière, A et K des paramètres indépendants du champ E et enfin $B = K.(\frac{\phi_0}{q})^{3/2}$

3.2.1.6 Mécanismes extrinsèques

Nous appelons mécanismes extrinsèques, les mécanismes additionnels causés par une dégradation du diélectrique ou par une mauvaise qualité de fabrication. Les mécanismes extrinsèques qui sont à l'origine de la perte de charges dans les mémoires non-volatiles à grille flottante sont au nombre de deux :

Les défauts des diélectriques

Les défauts qui existent dans le diélectrique qui entoure la grille flottante (oxyde de grille et oxyde interpoly) peuvent créer des chemins conducteurs qui vont décharger la grille flottante. Généralement, la conduction due aux défauts de l'oxyde de grille se fait par saut (ou "hopping"), avec une énergie d'activation de $0,6eV$ [Shiner'80] alors que la conduction due aux défauts dans l'oxyde interpoly se fait par effet "Poole-Frenkel" [Mielke'83]. On parle d'effet "Poole-Frenkel" lorsque la conduction est assistée par un seul piège et d'effet "Poole" lorsqu'elle a lieu avec plusieurs pièges. Tous ces mécanismes sont schématisés sur la figure 3.7.

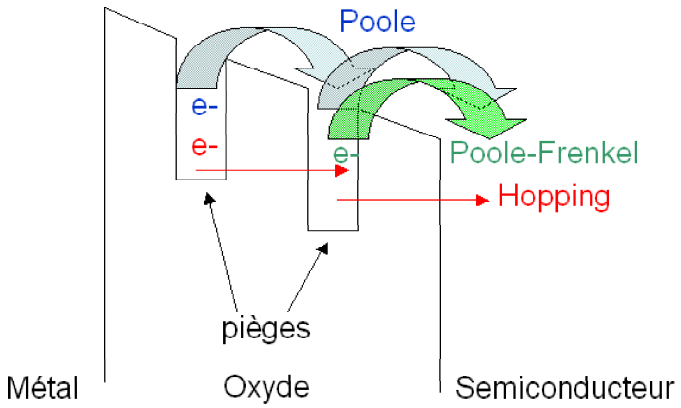


FIG. 3.7 – Effets "Poole", "Poole-Frenkel" et "Hopping"

La contamination ionique

La présence de contaminants dans la cellule peut modifier la tension de seuil ou provoquer des défaillances. Les ions contaminants sont la plupart du temps de charge positive (surtout des ions Sodium Na^+). Attirés par la grille flottante à cause du champ induit par la charge stockée, ces ions font écran à la charge emmagasinée, provoquant ainsi une "perte de charge virtuelle", mais ils créent aussi de véritables chemins de conduction provoquant des pertes de charges.

3.2.2 Protocole expérimental

Pour vérifier la capacité en rétention des mémoires, des normes JEDEC fixent des conditions standard de rétention [Jedec'04]. Celles-ci restent néanmoins très générales et d'après B. De Salvo elles ne sont pas fondées physiquement [DeSalvo'99]. Nous proposons donc une étude afin de mieux connaître le comportement de notre mémoire.

Nous avons donc choisi d'utiliser parmi ces standards les quatre températures suivantes : 25°C (afin de mesurer les pertes de charges à température ambiante), 85°C, 125°C et 150°C sur une durée de mille heures.

Les quatre wafers préalablement cyclés à 100.000 cycles selon les conditions définies précédemment ont ensuite été programmés aux quatre niveaux de tensions de seuil (00), (01), (10) et (11) sur chacun des wafers, puis mis à l'étuve aux quatre températures définies précédemment, selon le protocole résumé dans le tableau 3.5.

Wafers	Température	Nombre de Cellules au niveau(11)	Nombre de Cellules au niveau(10)	Nombre de Cellules au niveau(01)	Nombre de Cellules au niveau(00)
1	25°C	7	7	7	12
2	85°C	6	6	6	12
3	125°C	6	6	6	12
4	150°C	6	6	7	11

Tab. 3.5 – Récapitulatif du protocole expérimental

Des études préalables ont montré la possibilité de présence de charges piégées dans l'oxyde. Afin de vérifier cette hypothèse, nous avons conduit sur le wafer à 125°C une expérience d'effacement UV qui consiste à exposer le wafer à un rayonnement UV pendant quelques heures. On évacue ainsi les charges stockées pour retourner au niveau de V_T naturel. La température de 125°C a été choisie pour mener cette étude

du fait que nous voulions accélérer le plus fortement possible la perte de charge lors de la mesure en rétention, ce qui conduit à utiliser une température élevée, tout en se trouvant en dessous du seuil de recuit des défauts, généralement fixé entre 150°C et 200°C, et qu'ainsi les éventuelles différences observées seront dues à l'effacement UV et pas à un recuit des défauts qui améliorerait les caractéristiques [Naruke'88].

Dans la suite de ce manuscrit, nous utiliserons le terme "perte de charges" lorsque nous observerons la diminution d'une tension de seuil car ne pouvant pas accéder directement au nombre de charges stockées dans la grille flottante, nous en mesurerons une image par la mesure de la tension de seuil qui suit l'évolution de la charge par l'équation :

$$V_T = V_{T0} - \frac{Q_{FG}}{C_{pp}} \tag{3.3}$$

où V_{T0} est la tension de seuil de la cellule vierge, Q_{FG} la quantité de charges dans la grille flottante et $C_{pp} = \frac{\varepsilon_{ox}.\varepsilon_0.S_{pp}}{t_{pp}}$ la capacité interpoly où ε_{ox} est la permittivité diélectrique relative de l'oxyde utilisé, ε_0 est la permittivité diélectrique du vide, S_{pp} la surface de la capacité et t_{pp} l'épaisseur de l'isolant séparant les deux grilles.

Ainsi, toute perte de charge ΔQ_{FG} se traduit par une diminution de la tension de seuil $\Delta V_T = \frac{\Delta Q_{FG}}{C_{pp}}$

3.2.3 Présentation des résultats et Interprétations

A partir du protocole expérimental défini précédemment, nous avons effectué des lectures régulières des différentes tensions de seuil caractéristiques des quatre niveaux en fonction du temps de rétention afin de visualiser la dérive de ces tensions de seuil, traduisant les pertes de charges au cours du temps. A partir de ces pertes de charges, nous pourrons extrapoler les durées de rétention de l'information de chacun des quatre états.

3.2.3.1 Wafer W1 (25°C)

La figure 3.8 représente l'évolution sur mille heures des quatre niveaux de tension de seuil lors d'une rétention à température ambiante, soit 25°C.

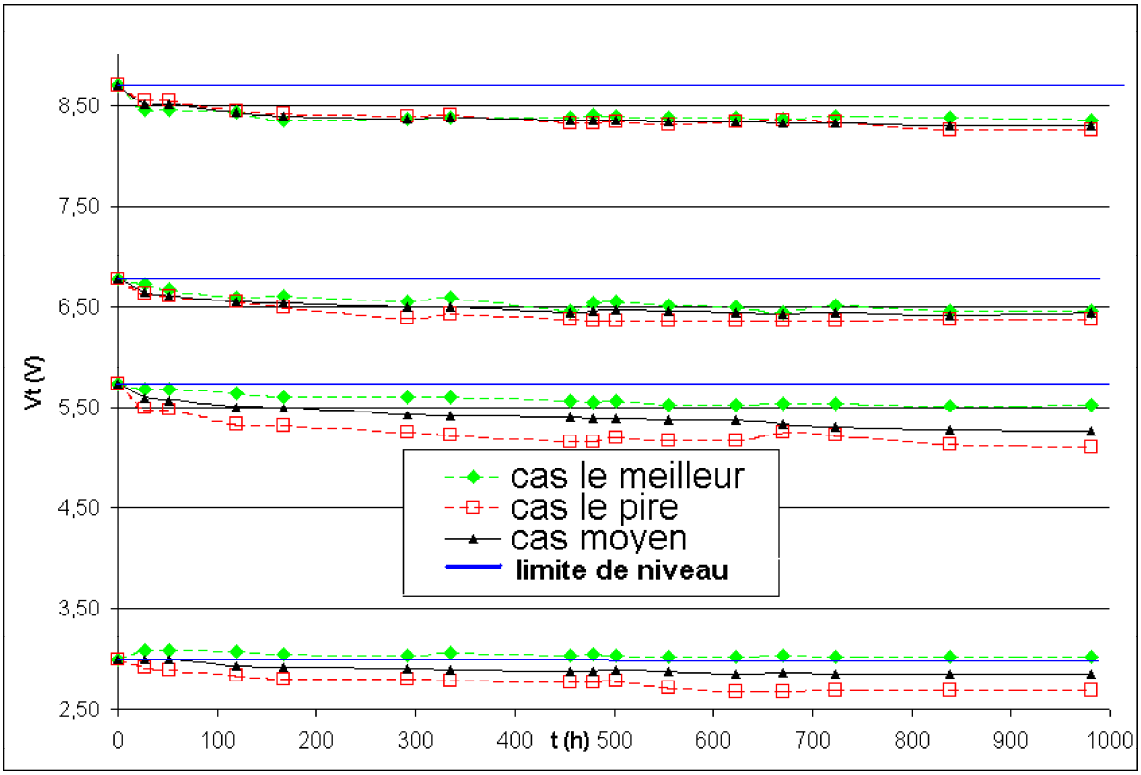


FIG. 3.8 – Récapitulatif de la rétention à 25°C

Il est possible de calculer les pentes de pertes de charges après les mille heures de rétention à 25°C afin d'observer la cinétique en régime permanent des pertes de charges pour chacun des niveaux logiques. Ces pentes géométriques, prises en échelle logarithmique du temps, sont résumées dans le tableau 3.6.

Niveau	Pentes de pertes de charges (V/décade)
(00)	0,0625
(01)	0,0626
(10)	0,0909
(11)	0,0437

TAB. 3.6 – Pentes à 1000h pour T=25°C

3.2.3.2 Wafer W2 (85°C)

La figure 3.9 représente l'évolution sur mille heures des quatre niveaux de tension de seuil lors d'une rétention à 85°C.

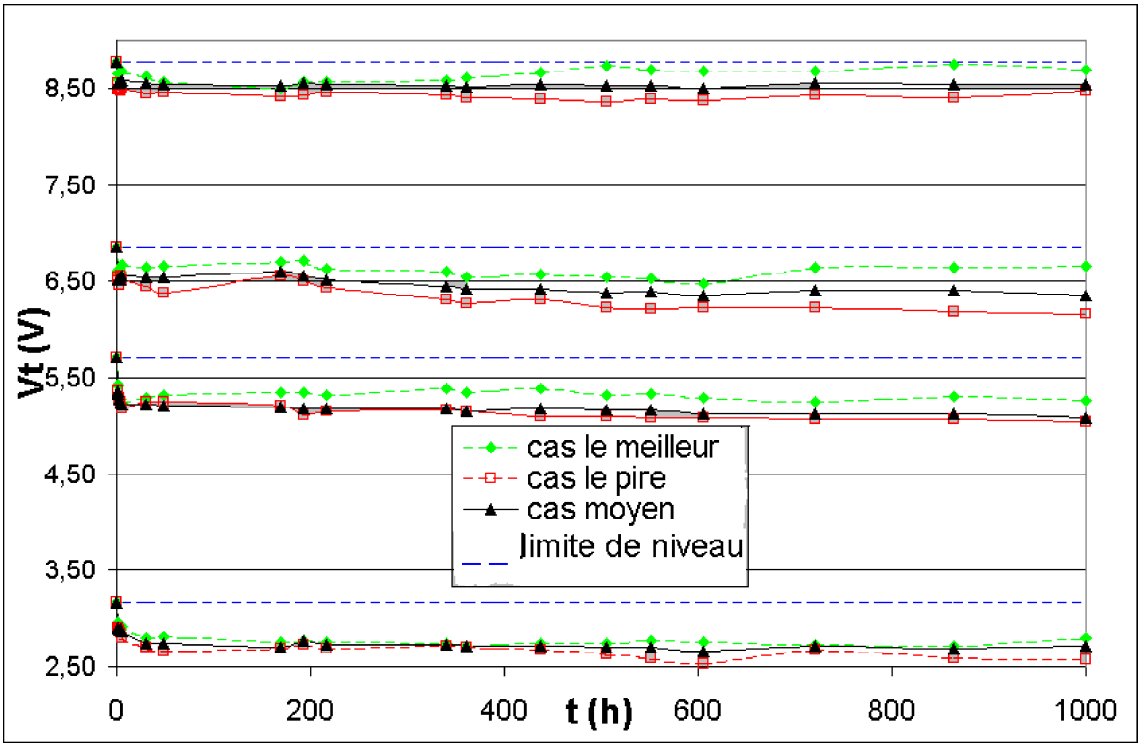


FIG. 3.9 – Récapitulatif de la rétention à 85°C

Comme nous l'avons fait pour $T = 25°C$ nous pouvons calculer les pentes de pertes de charges après les mille heures de rétention à 85°C afin d'observer la cinétique en régime permanent des pertes de charges pour chacun des niveaux logiques. Ces pentes géométriques, prises en échelle logarithmique du temps, sont résumées dans le tableau 3.7.

Niveau	Pentes de pertes de charges (V/décade)
(00)	0,0051
(01)	0,1233
(10)	0,0281
(11)	0,0334

TAB. 3.7 – Pentes à 1000h pour T=85°C

3.2.3.3 Wafer W3 (125°C)

La figure 3.10 représente l'évolution sur mille heures des quatre niveaux de tension de seuil lors d'une rétention à 125°C.

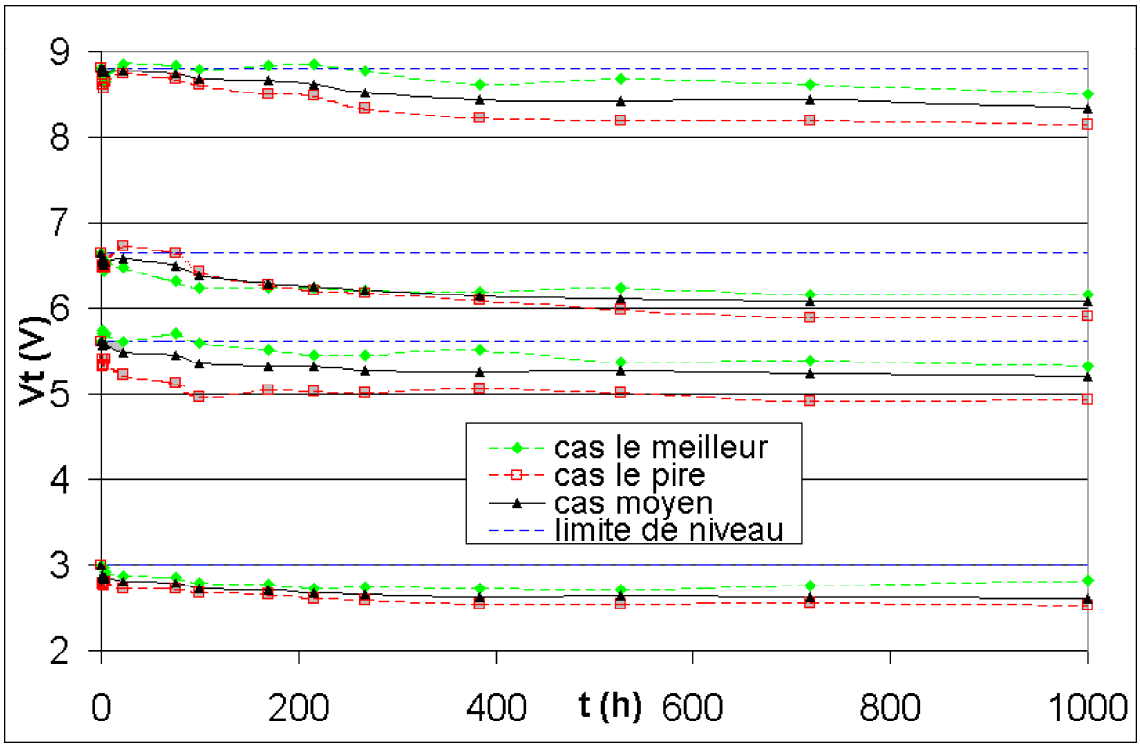


FIG. 3.10 – Récapitulatif de la rétention à 125°C

Nous pouvons également calculer les pentes de pertes de charges après les mille heures de rétention à 125°C afin d'observer la cinétique en régime permanent des pertes de charges pour chacun des niveaux logiques. Ces pentes géométriques, prises en échelle logarithmique du temps, sont résumées dans le tableau 3.8.

Niveau	Pentes de pertes de charges (V/décade)
(00)	0,1547
(01)	0,1396
(10)	0,0621
(11)	0,0517

TAB. 3.8 – Pentes à 1000h pour T=125°C

3.2.3.4 Wafer W4 (150°C)

La figure 3.11 représente l'évolution sur mille heures des quatre niveaux de tension de seuil lors d'une rétention à 150°C.

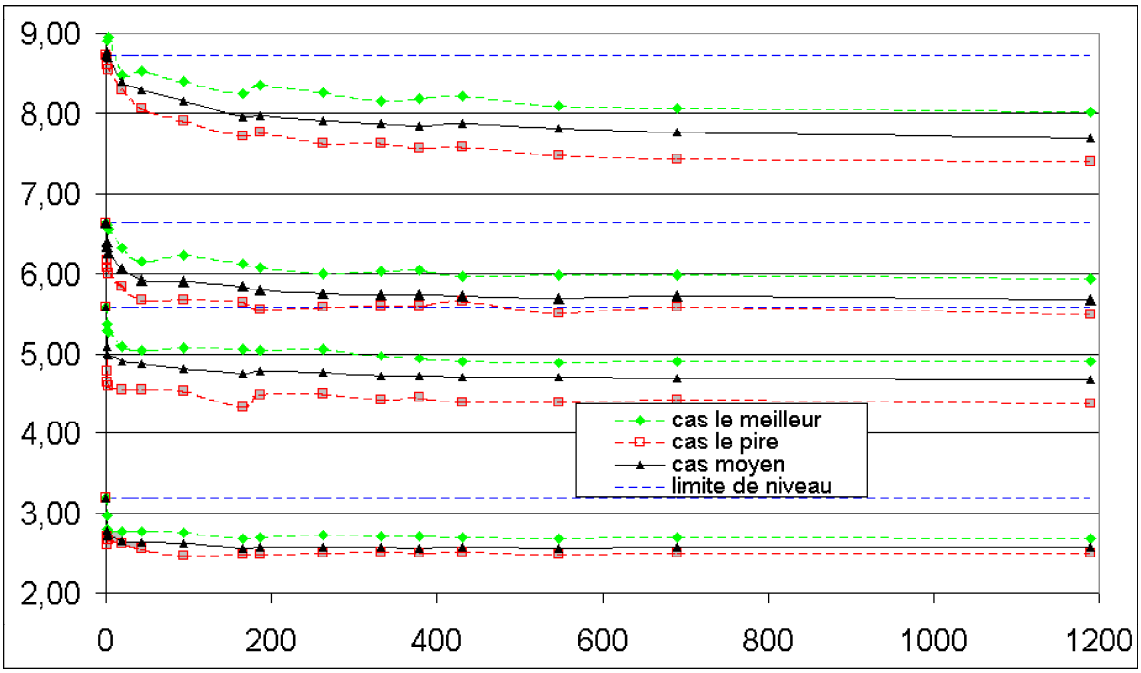


FIG. 3.11 – Récapitulatif de la rétention à 150°C

Nous pouvons également calculer les pentes de pertes de charges après les mille heures de rétention à 150°C afin d'observer la cinétique en régime permanent des pertes de charges pour chacun des niveaux logiques. Ces pentes géométriques, prises en échelle logarithmique du temps, sont résumées dans le tableau 3.9.

Niveau	Pentes de pertes de charges (V/décade)
(00)	0,1617
(01)	0,1081
(10)	0,0577
(11)	0,0307

TAB. 3.9 – Pentes à 1000h pour T=150°C

3.2.3.5 Conclusion sur les courbes de tenue en rétention

Nous pouvons remarquer sur toutes les courbes (hormis pour le wafer 3 à 125°C qui sera traité à part du fait de ses spécificités dans le protocole expérimental) une

première baisse assez brutale des niveaux de V_T, sur les 200 premières heures de rétention. Vient ensuite une perte plus faible et régulière dont la pente augmente avec le niveau de V_T. Ceci s'explique très bien par le fait que les pertes soient plus importantes lorsque le champ électrique est élevé. Il est également possible de remarquer que les pentes de pertes de charges à 25°C sont supérieures à celles à 85°C, ce qui semble très paradoxal. Cependant, le décrochement se produisant dans les 5 premières heures à 85°C induit des pertes qui sont approximativement les mêmes que celles des 100 premières heures à 25°C. Cela pourrait signifier que le même mécanisme est responsable de ces deux pertes mais qu'il est très fortement accéléré à la température de 85°C. Cette baisse est probablement due à la perte de charges qui ont été piégées au cours du cyclage. En effet, le cyclage sur 100.000 cycles a été réalisé en seulement 10 minutes, donc avec des temps de relaxation entre les cycles assez faibles (quelque μs). Cette diminution initiale ne dépend également pas du champ électrique car elle est présente avec la même amplitude pour chacun des 4 niveaux étudiés. Cet effet transitoire est donc activé thermiquement et sa durée dépend directement de la température. A 150°C, la même perte initiale apparaît dans les toutes premières heures de la rétention puis la courbe rejoint une perte plus classique. Si l'on observe attentivement les courbes à 150°C, on peut voir qu'il existe un mélange entre les niveaux (01) et (10). Cependant, ce mélange est uniquement dû au cas le pire qui correspond en fait à une unique cellule située sur le bord du wafer[4]. Pour toutes les autres cellules, le mélange n'apparaît pas.

3.2.3.6 Cas particulier du Wafer W3 à 125°C

Dans le cas du wafer **W3** à 125°C, il convient de rappeler qu'un effacement préalable par rayonnement UV a été effectué afin de dépiéger d'éventuelles charges emmagasinées au cours du cyclage. On observe alors que la perte initiale très importante sur les autres wafers est cette fois-ci très limitée, voire quasi-inexistante. Cela conforte l'hypothèse expliquant la première partie de nos courbes en rétention, selon laquelle les charges piégées lors du cyclage se dépiègent lors des premières heures de rétention.

3.3 Modélisation

L'objectif de cette étude était de comprendre les phénomènes physiques régissant les pertes de charges dans la cellule mémoire. Pour cela nous devons identifier le mécanisme prépondérant en comparant les mesures aux différents modèles existants, en vue d'en déduire le mécanisme mis en jeu. Des études précédentes ont montré qu'après un temps suffisamment long, on atteint une saturation de la perte de charges qui peut être modélisée par une équation de type Fowler-Nordheim [Kameyama'00].

[4]Au cours du process de fabrication, il arrive que les bords du wafers ne subissent pas exactement les mêmes conditions que le reste du wafer et que les propriétés des cellules en périphérie se trouvent dégradées, ce qui semble être le cas ici.

Auparavant, un autre phénomène serait responsable de la perte importante de charges, de type dépiégeage des charges stockées dans l'oxyde au cours du cyclage, ce que nous tenterons de vérifier dans la suite de ce chapitre.

3.3.1 Détermination du nombre de phénomènes physiques mis en cause

De nombreuses études ont démontré que tous les mécanismes de fuite pouvaient être modélisés par des équations simples de type Fowler-Nordheim avec une hauteur de barrière éventuellement modifiée [Bhattacharyya'84][Kayemana'00][Ielmini'05] ou de type Poole-Frenkel [Cheng'87][Cheng'88] :

$$J_{FN} = A(T).E_{ox}^2.\exp(\frac{-B(T)}{E_{ox}}) \tag{3.4}$$

$$J_{PF} = A'(T).E_{ox}.\exp(q.\beta_{PF}.\frac{E_{ox}^{1/2}}{kT}) \tag{3.5}$$

Afin de mettre en évidence la sensibilité des mécanismes mis en jeu à la température de rétention, nous avons défini le facteur Γ selon la formule :

$$\Gamma_{(X°C)}(t) = \frac{\Delta V_{T_{(X°C)}}(t)}{\Delta V_{T_{(25°C)}}(t)} \tag{3.6}$$

La figure 3.12 représente ce facteur pour les niveaux (10) et (01) à la température 150°C et met en évidence deux mécanismes distincts de pertes de charges :

- un premier mécanisme fortement amplifié, d'un facteur 4 à 2.5 environ, par la température avant 200 heures de rétention ;
- un second mécanisme moins amplifié, d'un facteur 1.9 à 2.5 environ, par la température après 200 heures.

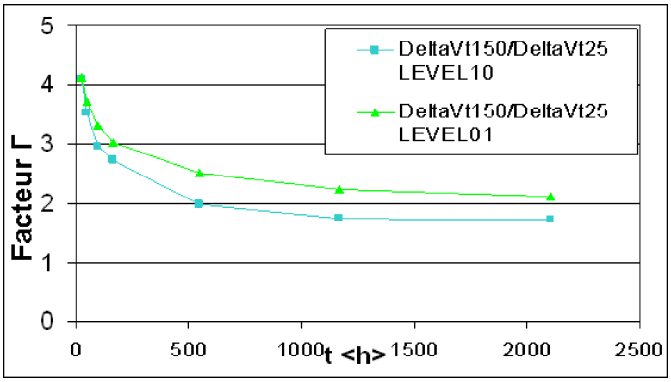


FIG. 3.12 – Facteur Γ à 150°C

Après avoir identifié la présence de deux mécanismes, nous pouvons maintenant tenter de modéliser chacun de ces deux mécanismes.

3.3.2 Modélisation des pertes après 200 heures

Si l'on s'intéresse dans un premier temps à la perte de charges après un temps suffisamment long, soit environ 200 heures, le phénomène est classiquement modélisé par une équation du type équation Fowler-Nordheim [Kameyama'00], ce que nous allons chercher à vérifier dans la suite de cette étude.

3.3.3 Extraction des paramètres de l'équation Fowler-Nordheim A et B

On peut chercher à déterminer les paramètres A et B intervenant dans l'équation du courant Fowler-Nordheim, dont la densité de courant a été donnée dans l'équation (3.2) : [FowlerNordheim'28]

$$I_{FN} = J_{FN}.S_{tun} = A.S_{tun}.E_{ox}^2 \exp(\frac{-B}{E_{ox}}) \tag{3.7}$$

avec S_{tun} la surface de l'oxyde tunnel.

Pour ce faire, nous utiliserons des courbes I-V tracées expérimentalement à partir d'une capacité de grande superficie. Le choix de travailler sur une telle capacité au lieu de la véritable cellule est due au fait que nous devons accéder directement à la grille flottante (ce qui est impossible sur la cellule mémoire) mais également que nous avons besoin d'une surface élevée pour obtenir des courants suffisamment grands pour être mesurés.

Nous déterminerons ensuite A et B en traçant nos courbes I-V dans le plan de Fowler-Nordheim , c'est-à-dire que l'on représentera $Y = ln(I_{FN}/S_{tun}E_{ox}^2) = f(1/E_{ox}) = f(X)$, ce qui doit nous permettre d'obtenir une droite d'équation :

$$Y = -B.X + ln(A) \tag{3.8}$$

La pente de la droite permet d'extraire directement le paramètre B, tandis que l'ordonnée à l'origine fournit le paramètre A.

3.3.4 Le modèle utilisé

Le modèle de perte de charges, représenté par la diminution de la tension de seuil $V_T(t)$, qui a été utilisé suit le cheminement suivant :

$$V_T(t) \overset{(3.9)}{\rightarrow} V_{fg}(t) \overset{(3.10)}{\rightarrow} E_{ox}(t) \overset{(3.11)}{\rightarrow} I_{fuite}(t) \overset{(3.12)}{\rightarrow} \Delta Q_{fg}(t+\Delta t) \overset{(3.13)}{\rightarrow} V_T(t+\Delta t)$$

et on reboucle en utilisant $V_T(t + \Delta t)$ en entrée.

Connaissant la valeur $V_T(t)$ de la tension de seuil à un instant t, il est possible d'en déduire la tension de la grille flottante par l'expression :

$$V_{fg}(t) = \alpha_g.V_T(t) \tag{3.9}$$

avec α_g le coefficient de couplage entre la grille de contrôle et la grille flottante valant ≈ 0.52.

Le champ électrique aux bornes de l'oxyde tunnel se calcule alors grâce à :

$$E_{ox}(t) = \frac{V_{fg}(t) - \alpha_g.V_{T0}}{t_{ox}} = \frac{\alpha_g(V_T(t) - V_{T0})}{t_{ox}} \tag{3.10}$$

avec $t_{ox} = 10.2nm$ l'épaisseur de l'oxyde tunnel et $V_{T0} = 3V$ la tension de seuil naturelle de la cellule.

Nous pouvons alors utiliser l'expression (3.7) pour obtenir la valeur du courant de fuite I_{fuite} à l'instant t :

$$I_{fuite}(t) = A.S_{ox}.E_{ox}(t)^2 \exp(\frac{-B}{E_{ox}(t)}) \tag{3.11}$$

En utilisant l'expression du courant comme étant la dérivée de la charge par rapport au temps $I = \frac{dQ}{dt}$, cela donne pour la charge de la grille flottante Q_{fg} :

$$\Delta Q_{fg}(t + \Delta t) = I_{fuite}(t).\Delta t \tag{3.12}$$

D'après l'équation (3.3), la tension de seuil peut être reliée directement à la charge et par conséquent la variation de la tension de seuil dans un intervalle Δt peut être directement reliée à la variation de charge pendant ce même intervalle de temps Δt :

$$V_T(t + \Delta t) = V_T(t) - \frac{\Delta Q_{fg}(t + \Delta t)}{C_{pp}} \tag{3.13}$$

avec C_{pp} la capacité entre la grille de contrôle et la grille flottante, valant $\frac{\epsilon_0.\epsilon_{ONO}.S_{pp}}{t_{pp}}$

3.3.5 Résultats de la modélisation – Comparaison à l'expérience

Les courbes de pertes de charges des quatre niveaux sont tracées dans le plan de Fowler-Nordheim (cf. figures 3.13 à 3.16), c'est-à-dire que l'on représente $ln(J/SE^2) = f(1/E)$, après 200 heures. Ainsi, si le mécanisme responsable de la fuite de charge suit l'équation Fowler-Nordheim (3.7), nous devons observer une droite dans ce système d'axes.

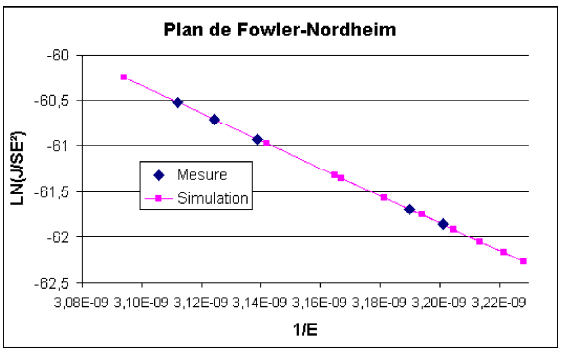


FIG. 3.13 – Modélisation des pertes de charges du niveau (00) à 150°C

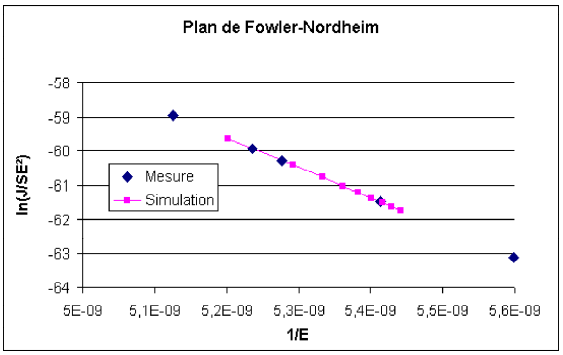


FIG. 3.14 – Modélisation des pertes de charges du niveau (01) à 150°C

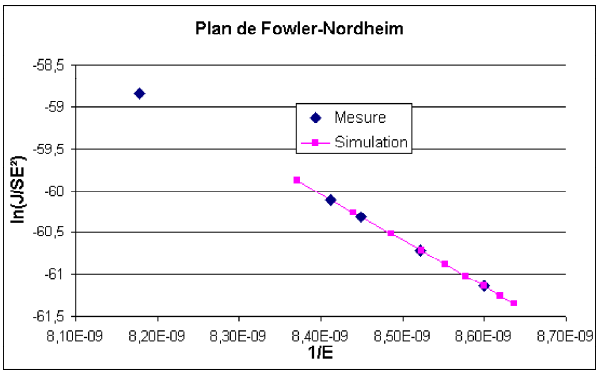


FIG. 3.15 – Modélisation des pertes de charges du niveau (10) à 150°C

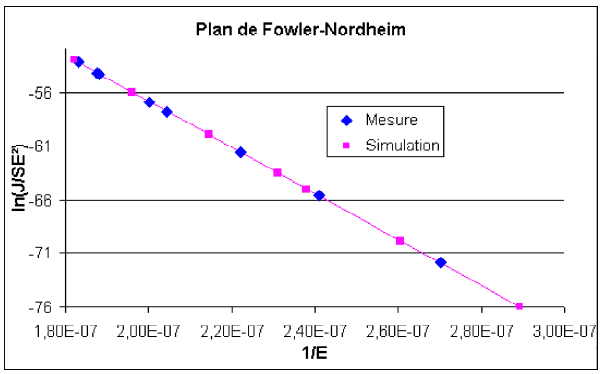


FIG. 3.16 – Modélisation des pertes de charges du niveau (11) à 150°C

Nous pouvons remarquer que le modèle proposé correspond bien aux mesures pour les quatre niveaux (00), (01), (10) et (11). Le fait que l'on obtienne une droite dans le plan de Fowler-Nordheim semble valider notre hypothèse de fuite par un courant modélisable par l'équation Fowler-Nordheim après 200 heures de rétention à 150°C. Nous devons néanmoins extraire les valeurs de hauteurs de barrière φ_0 pour identifier un véritable courant Fowler-Nordheim, si la hauteur de barrière obtenue correspond à la véritable hauteur de barrière Si-SiO2 de l'ordre de 3.2V, ou identifier un mécanisme autre qu'un courant Fowler-Nordheim mais modélisable par l'équation Fowler-Nordheim, si la hauteur de barrière extraite est inférieure à 3.2V.

Ayant extrait les paramètres A et B dans le paragraphe 3.3.3, nous pouvons inverser les équations (3.14) et (3.15), caractéristiques des paramètres Fowler-Nordheim, pour remonter à la hauteur de barrière φ_0 par les équations (3.16) et (3.17).

$$A = \frac{m}{m^*}.\frac{q^2}{8\pi.h.\varphi_0} \quad (3.14)$$

$$B = \frac{8\pi}{3h}\sqrt{2qm^*\varphi_0^3} \quad (3.15)$$

$$\varphi_0(A) = \frac{m}{m^*}.\frac{q^2}{8\pi.h.A} \quad (3.16)$$

$$\varphi_0(B) = (\frac{3.h.B}{8\pi.\sqrt{2.q.m^*}})^{2/3} \quad (3.17)$$

Pour calculer φ_0, nous devons fixer la valeur de la masse effective m^* des électrons dans le SiO2, considérée indépendante de la température [Pananakakis'95]. Plusieurs valeurs cohabitent dans la littérature mais la valeur la plus fréquemment utilisée est $m^* = 0.5m = 0.5 \times 9.11e-31kg = 4.555e-31kg$ [Weinberg'82][Fischetti'87][Suné'87] [Yang'94][Pananakakis'95]. A partir de ces valeurs de m^*, A et B, il ne nous est pas possible d'extraire des hauteurs de barrière φ_0 communes selon que nous utilisons

le paramètre A ou B. Cela traduit le fait que même si nous pouvons modéliser le courant de fuite par une équation de type Fowler-Nordheim, il ne s'agit en aucun cas d'un véritable courant Fowler-Nordheim, comme nous pouvions le supposer depuis le départ du fait de la nature même de ce courant, typique des phases d'injection au cours desquelles le champ électrique est élevé.

3.3.6 Relation entre les températures

Suite à ces mesures de rétention, nous pouvons chercher des relations entre les pertes aux différentes températures. Nous nous fixons pour cela une mesure, extrapolée si nous n'avons pas atteint la critère après mille heures, des temps de rétention à 15% de perte de charges, soit correspondant à un $DeltaV_T = 0,85 \times \Delta V_{Tinitial}$. L'ensemble de ces temps de rétention sont synthétisés dans le tableau 3.10.

Niveau	T=25°C	T=85°C	T=125°C	T=150°C
(00)	1,50E+10	-	252926	6518
(01)	4,79E+08	6,26E+04	16850	975
(10)	2,59E+11	1,42E+08	1,21E+07	238
(11)	897254	653	2879	4

Tab. 3.10 – Temps de rétention à 15% en heures

La loi classique de pertes en rétention utilisée est la loi d'Arrhenius en 1/T mais des études ont montré qu'une loi en T pouvait être plus appropriée [DeSalvo'99]. Nous avons donc essayé de représenter nos relations entre températures selon des lois en 1/T et en T afin de vérifier l'une ou l'autre des hypothèses. Il s'avère finalement d'après les figures 3.17 à 3.20 que la loi en T donne de meilleurs résultats, comme démontré par B. DeSalvo.

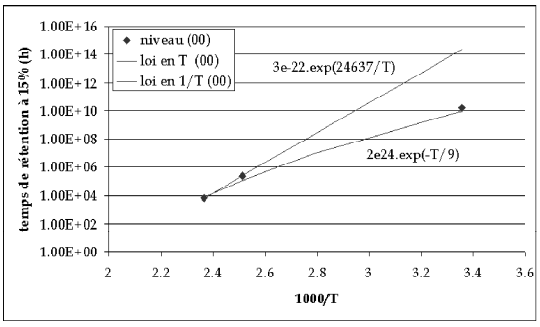


Fig. 3.17 – Modélisation en 1/T et T des pertes de charges du niveau (00)

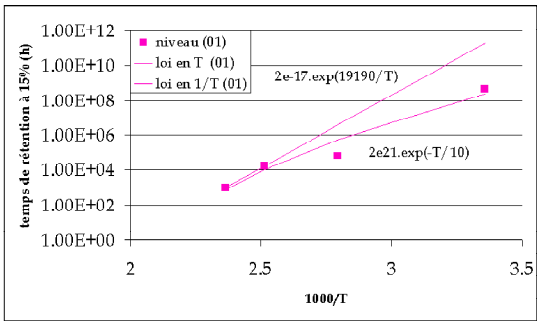


Fig. 3.18 – Modélisation en 1/T et T des pertes de charges du niveau (01)

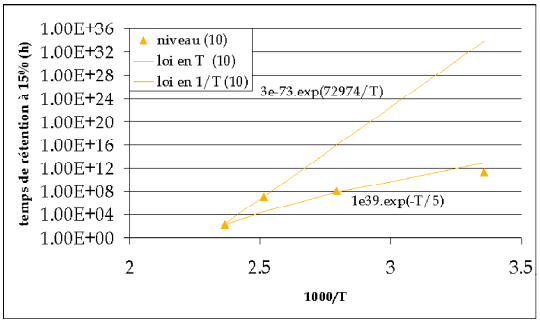


FIG. 3.19 – Modélisation en 1/T et T des pertes de charges du niveau (10)

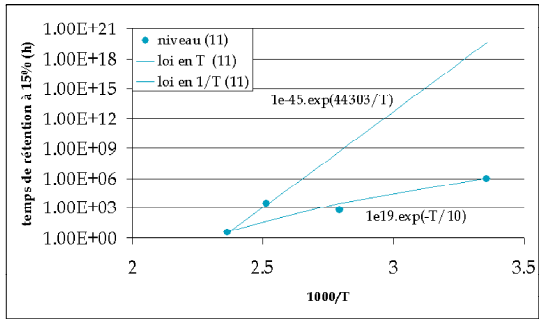


FIG. 3.20 – Modélisation en 1/T et T des pertes de charges du niveau (11)

Si l'on utilise la formule (3.18) de la loi en T reliant le temps de rétention t_R à la température T, fournie par DeSalvo, nous trouvons comme une température caractéristique de la rétention de donnée $T0_{DR}$ comprise entre 5 et $10K$, ce qui est du même ordre de grandeur que la valeur donnée par DeSalvo $T0_{DR} = 21K$.

$$t_R = t_0.e^{-(T/T0_{DR})} \tag{3.18}$$

où t_0 est le temps de rétention à $T = 0K$

Nous voyons que nous avons mis en place une extrapolation fiable du temps de rétention de nos cellules pour les quatre niveaux considérés, à toutes les températures de 25°C à 150°C.

3.3.7 Modélisation de la perte initiale

Après 200 heures de rétention, nous avons réussi à identifier un mécanisme décrit par une équation de type Fowler-Nordheim comme étant responsable des pertes de charges. Cependant, reste à expliquer la perte initiale importante.

Le mécanisme identifié après les 200 premières heures de rétention existe cependant depuis le début de la rétention, c'est pourquoi nous allons tout d'abord l'étendre à l'ensemble de la mesure de la perte de charges puis lui superposer un second mécanisme si nécessaire.

Nous voyons bien sur la figure 3.21 que le mécanisme que nous avons identifié après 200 heures de rétention ne peut pas être le seul responsable de la perte totale de charges. Un premier mécanisme, prépondérant avant 200 heures, est bien présent, causant la forte perte initiale de charges, comme nous l'avions mis en évidence grâce au calcul du facteur Γ en température au paragraphe 3.3.1. Nous pouvons tenter de superposer soit un second mécanisme modélisé par une équation de type Fowler-Nordheim (Fig. 3.22), soit un mécanisme modélisé par une équation de type Poole-Frenkel (Fig. 3.23).

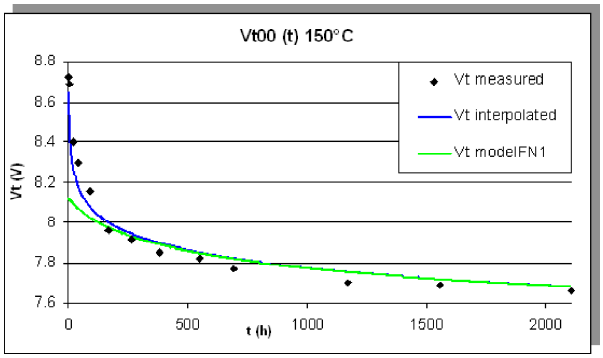


FIG. 3.21 – Modélisation de la rétention du niveau 00 due au mécanisme de saturation Fowler-Nordheim

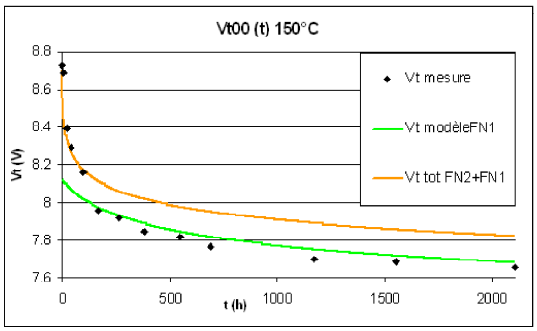


FIG. 3.22 – Modélisation de la rétention du niveau 00 due au mécanisme de saturation Fowler-Nordheim, combiné à un second mécanisme Fowler-Nordheim

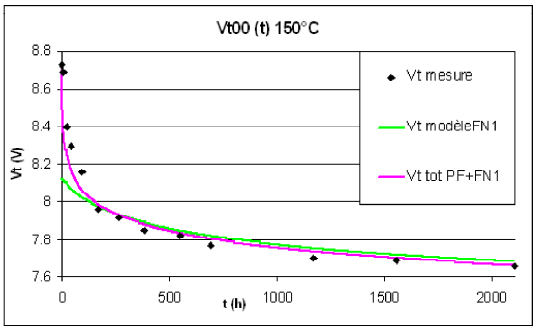


FIG. 3.23 – Modélisation de la rétention du niveau 00 due au mécanisme de saturation Fowler-Nordheim, combiné à un mécanisme Poole-Frenkel

De façon évidente, l'ajout d'un mécanisme Poole-Frenkel permet une bien meilleure modélisation de la rétention, quel que soit le temps pris en compte.

En réalisant cette étude sur l'ensemble des trois wafers placés à 25°C, 85°C et 150°C, nous avons pu identifier à chaque fois un premier mécanisme modélisable par l'équation Poole-Frenkel, principalement responsable de la perte apparaissant lors des 200 premières heures, superposé à un mécanisme modélisable par l'équation Fowler-Nordheim, principalement responsable de la perte après 200 heures, bien que présent dès les premières heures de rétention.

En ce qui concerne le wafer placé à 125°C qui, rappelons le, a subi une exposition aux UV afin de dépiéger les charges emmagasinées lors du cyclage, nous avons pu modéliser sa perte totale de charges par un seul mécanisme de type Fowler-Nordheim, valable quel que soit le temps de rétention. Cela conforte donc notre hypothèse que le premier mécanisme prépondérant, responsable de la perte rapide de charges, soit directement lié aux pièges créés lors du cyclage préalable. Une fois que ceux-ci ont

été dépiégés dans les premières heures de rétention la perte de charge se fait par un mécanisme de type courant tunnel, avec des pertes beaucoup moins grandes. Le wafer à 125°C ayant subi l'effacement UV avant la mesure en rétention avait déjà perdu tout ou grande partie de ses pièges à t=0. Le seul mécanisme de fuite restant était donc le mécanisme de fuite de type courant tunnel qui s'applique donc parfaitement tout au long de la mesure de perte de charges, ce que nous pouvons résumer dans le tableau 3.11 :

T	niveau (00)		niveau (01)		niveau (10)		niveau (11)	
(°C)	$t < 200h$	$t > 200h$	$t < 200h$	$t > 200h$	$t < 200h$	$t > 200h$	$t < 200h$	$t > 200h$
25	PF	FN	PF	FN	PF	FN	PF	FN
50	PF	FN	PF	FN	PF	FN	PF	FN
125	FN	FN	FN	FN	FN	FN	FN	FN
150	PF	FN	PF	FN	PF	FN	PF	FN

TAB. 3.11 – Résumé des types d'équations permettant la modélisation des mécanismes responsables des pertes de charges

3.3.8 Modélisation d'un gain de charges au cours des premières heures de rétention

Dans certains cas, nous avons pu observer une augmentation de la tension de seuil de nos cellules (cf. figure 3.24), ce qui est contraire à la théorie de pertes de charges à travers l'oxyde tunnel mais qui a également été reporté dans la littérature [Kameyama'00].

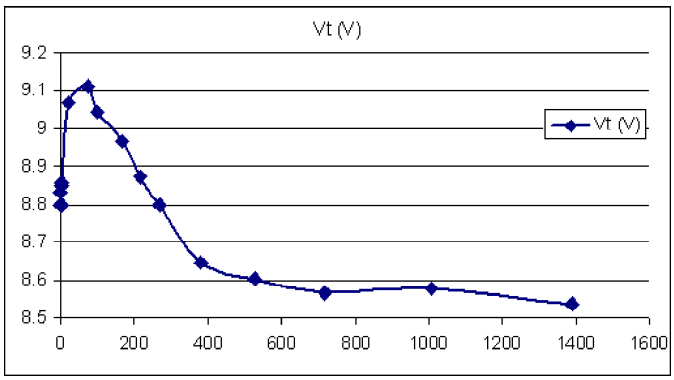


FIG. 3.24 – Observation d'un gain de charges lors des premières heures de rétention

Ce phénomène peut cependant trouver une explication dans un déplacement des charges piégées dans l'oxyde tunnel au cours du cyclage. [Dumin'93]

D'après Dumin, si l'on considère un front plan de charges piégées dans l'oxyde, son déplacement dans cet oxyde est régi par l'équation (3.19) :

$$x(t) = (2\beta)^{-1}.\ln(\frac{t-\tau}{t_0}) \tag{3.19}$$

où β est une constante, τ est le retard d'activation du déplacement du front de charges et t_0 une constante de temps caractéristique.

En utilisant cette équation avec un front de charges situé au milieu de l'oxyde tunnel à t=0 et se déplaçant progressivement vers le substrat, nous obtenons les simulations de tensions de seuil des figures 3.25 et 3.26 avec $\beta = 30$, un retard d'activation $\tau = 4h$ et $t_0 = 1e-10s$:

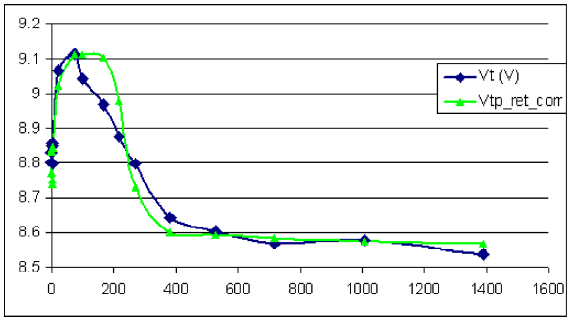


FIG. 3.25 – Modélisation du gain de charges observé lors des premières heures de rétention en échelle linéaire

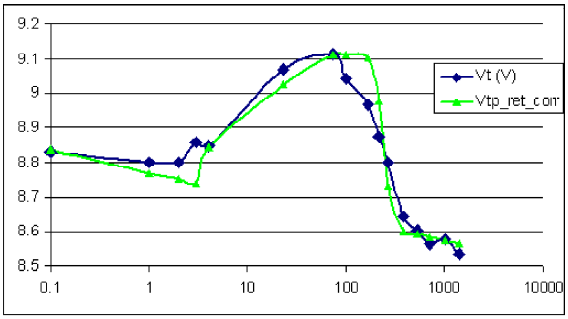


FIG. 3.26 – Modélisation du gain de charges observé lors des premières heures de rétention en échelle logarithmique

Ce modèle de déplacement d'un front de charges à l'intérieur de l'oxyde tunnel lors des premières heures de la rétention reproduit très bien les mesures effectuées.

3.4 Conclusion

Au cours de ce chapitre, nous avons proposé une modélisation de pertes de charges en rétention après cyclage de cellules Flash en architecture NOR multi-niveaux. Deux bits au lieu d'un seul étaient en effet stockés sur une seule cellule, ce qui a justifié cette étude de discrimination des états logiques. Nous avons pu mettre en évidence pour les quatre températures étudiées un premier dépiégeage de charges lors des 200 premières heures de rétention, que nous avons modélisé par une équation de type Poole-Frenkel. A cette équation, nous avons superposé une équation de type Fowler-Nordheim qui modélise le phénomène prépondérant après la forte perte initiale. Nous avons également modélisé un "gain" fictif de charges, qui se produit parfois lors des mesures en rétention, par le déplacement d'un front de charges du milieu de l'oxyde tunnel vers le substrat.

Références bibliographiques du chapitre 3

[FowlerNordheim'28] R.H. Fowler, L. Nordheim
"Electron Emission in intense electric fields"
Proc. Soc. London Ser., A119, 781, pp.173-181, 1928.

[Shiner'80] R.E. Shiner, J.M. Caywood and B.L. Euzent
"Data retention in EPROMs"
Proceedings of IRPS, pp.238, 1980.

[Nozawa'82] H. Nozawa, S. Kohyama
"A thermionic electron emission model for charge retention in SAMOS structure"
Japanese Journal of Applied Physics, 21, L111, 1982.

[Weinberg'82] Z.A. Weinberg
"On tunneling in metal-oxide-silicon structures"
Journal of Applied Physics, vol.53, no.7, pp.5052-5056, 1982.

[Mielke'83] N.R.Mielke
"New EPROM data-loss mechanisms"
Proceedings of IRPS, pp.106, 1983.

[Bhattacharyya'84] A. Bhattacharyya
"Modeling of write/erase and charge retention characteristics of floating gate EEPROM devices"
Solid-State Electronics, vol.27, no.10, pp.899-906, 1984.

[Cheng'87] X.R. Cheng, B.Y. Liu, Y.C. Cheng
"Electrical conduction in thin thermally nitrided SiO2"
Applied Surface Science, 30, pp.237-243, 1987.

[Fischetti'87] M.V. Fischetti, D.J. DiMaria, L. Dori, J. Batey, E. Tierney, J. Stasiak
"Ballistic electron transport in thin silicon dioxide films"
Physical Review B, vol.35, no.9, pp.4404-4415, 1987.

[Suné'87] J. Suné, Y. Placencia, F. Campabadal, X. Aymerich
"Characterization of metal-SiO2-Si interface roughness by electrical methods"
Surface Science, vol.189-190, pp.346-352, 1987.

[Cheng'88] X.R. Cheng, Y.C. Cheng, B.Y. Liu
"Nitridation-enhanced conductivity behavior and current transport mechanism in thin thermally nitrided SiO2"
Journal of Applied Physics, vol.63, no.3, pp.797-802, 1988.

[Naruke'88] K. Naruke, S. Taguchi, M. Wada
"Stress Induced Leakage Current limiting to scale down EEPROM tunnel oxide thickness"
Proceedings of IEDM, pp.424-427, 1988.

[Crisenza'91] G. Crisenza, G. Ghidini, M. Tosi
"Floating Gate Memories reliability"
Proceedings of ESREF, pp.505, 1991.

[Mazoyer'92] P. Mazoyer
"Analyse et caractérisation des mécanismes de perte de charge relatifs aux diélectriques multicouches du point mémoire"
Thèse de doctorat, Université Joseph Fourier de Grenoble, 1992.

[Dumin'93] D.J. Dumin, J.R. Maddux, R.S. Scott, R. Subramoniam
"A model relating wearout induced physical changes in thin oxides to the statistical description of breakdown"
Proceedings of IRPS, pp.285, 1993.

[Yang'94] B.L. Yang, H. Wong, Y.C. Chen
"Study of process-dependent electron-trapping characteristics of thin nitrided oxides"
Solid-State Electronics, vol.37, no.3, pp.481-486, 1994.

[Papadas'95] C. Papadas, G. Pananakakis, G. Ghibaudo, C. Riva, F. Pio, P. Ghezzi
"Modeling of the intrinsec retention characteristics of Flotox EEPROM cells under elevated temperature conditions"
IEEE Transactions on Electron Devices, vol.42, no.4, pp.678, 1995.

[Pananakakis'95] G. Pananakakis, G. Ghibaudo, R. Kies, C. Papadas
"Temperature dependence of the Fowler-Nordheim current in Metal-Oxide-Degenerate Semiconductor structures"
Journal of Applied Physics, vol.78, no.4, pp.235, 1995.

[Candelier'97] P. Candelier
"Contribution à l'amélioration de la fiabilité des mémoires non-volatiles de type Flash-EEPROM"
Thèse de doctorat, Université Joseph Fourier de Grenoble, 1997.

[DeSalvo'99] B. De Salvo, G. Ghibaudo, G. Pananakakis, G. Reimbold, F. Mondond, B. Guillaumond, P. Candelier
"Experimental and theoretical investigation of nonvolatile memory data retention"
IEEE Transactions on Electron Devices, vol.46, no.7, pp.1518, 1999.

[Kameyama'00] H. Kameyama et al.
"A new data retention mechanism after endurance stress on flash memory"
Proceedings of IRPS, 2000.

[Ielmini'01] D. Ielmini, A.S. Spinelli, A.L. Lacaita, L. Confalonieri, A. Visconti
"New technique for fast characterization of SILC distribution in Flash arrays"
Proceedings of IRPS, pp.73-80, 2001. [Modelli'01] A. Modelli, A. Manstretta, G.

Torelli
"Basic Feasibility constraints for multilevel CHE-programmed Flash memories"
IEEE Transactions on Electron Devices, vol.48, no.9, pp.2032, 2001.

[SEMATECH'03] International SEMATECH
"Critical reliability challenges for the International Technology Roadmap for Semiconductors (ITRS)"
International SEMATECH Technology Transfer, 2003.

[Jedec'04] Extrait des Normes JEDEC
"High temperature storage life"
JESD22-A103C, 2004.

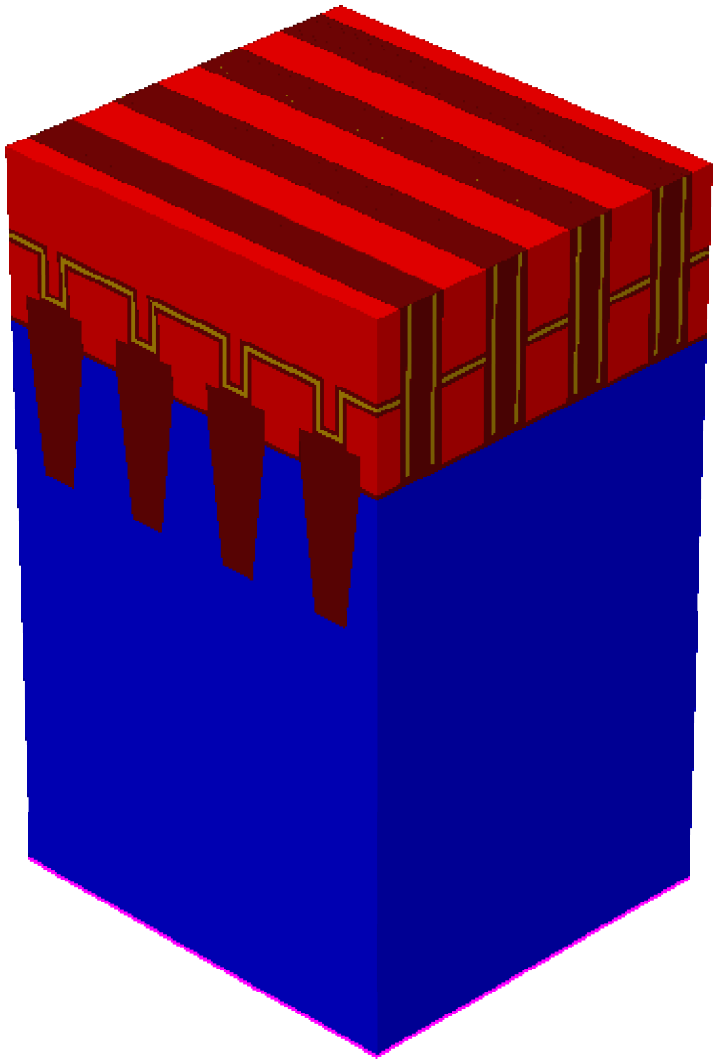
[Razafindramora'04] J.B. Razafindramora
"Modélisation et caractérisation de transistors MOS appliquées à l'étude de la programmation et du vieillissement de l'oxyde tunnel des mémoires EEPROM"
Thèse de doctorat, Université de Provence, 2004.

[ITRS'05] International Technology Roadmap for Semiconductors
"Process Integration, Devices and Structures"
ITRS Roadmap, 2005 Edition.

[Ielmini'05] D. Ielmini, A.S. Spinelli, A.L. Lacaita
"Recent developments on Flash memory reliability"
Microelectronic Engineering, 80, pp.321-328, 2005.

Chapitre 4

Perturbations



Sommaire

Ce chapitre a pour but d'étudier diverses perturbations pouvant intervenir sur les cellules mémoires et pouvant être causées soit par l'application de polarisations qui créent un effet non souhaité, soit par des couplages parasites entre cellule. Nous traiterons donc dans un premier temps des perturbations causées par une polarisation appliquée sur la grille de contrôle d'une cellule en architecture NOR, avant et après cyclage, puis sur la structure S16 en architecture NAND. Nous aborderons ensuite un autre aspect des perturbations avec l'étude de la dégradation des cellules inhibées au cours de laquelle nous exposerons les perturbations créées par les capacités parasites de couplage entre cellules à l'intérieur de la matrice mémoire. Nous développons pour cela des simulations TCAD bidimensionnelles et tridimensionnelles. Nous présenterons également une explication du mécanisme de dégradation des cellules inhibées lors du cyclage de la cellule sélectionnée en nous basant à la fois sur des simulations TCAD et sur des mesures sur silicium.

4.1 Evaluation des perturbations de grille sur cellules mémoires en architecture NOR

Au cours du développement d'une future technologie en architecture NOR par l'entreprise, un problème de perturbations dues à la polarisation de grille s'est posé, c'est pourquoi l'étude s'est orientée vers ce phénomène et que les mesures de disturb réalisées dans la suite de ce paragraphe sont des mesures de program disturb, perturbation décrite au paragraphe 1.2.4.1.

4.1.1 Avant cyclage

Les conditions de polarisations utilisées pour les expériences présentées dans ce chapitre sont décrites sur la figure 4.1. Les signaux sont des pulses carrés ayant tous la même cinétique.

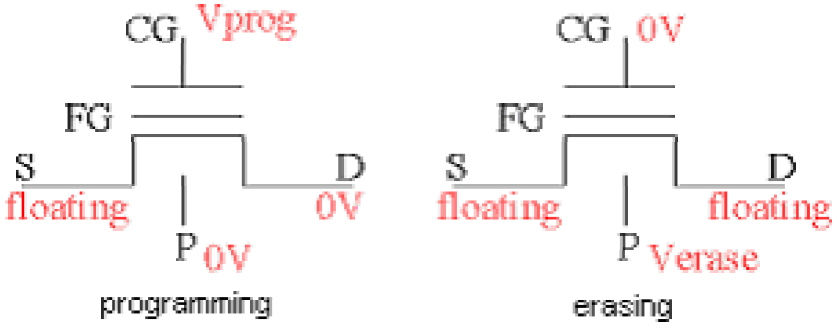


FIG. 4.1 – Conditions de programmation utilisées lors des mesures de disturb

Les niveaux "haut" et "bas" correspondront respectivement à des valeurs de tensions de seuil V_T de $+4V$ et $0V$.

4.1.1.1 Détermination de l'efficacité de programmation avant cyclage

Partant d'une cellule à l'état effacé, une série de mesures avec des tensions de programmation V_{prog} variant de $+10V$ à $+19V$ a été effectuée afin de montrer l'influence de la tension de grille de contrôle WL sur la tension de seuil V_T au cours du temps, reportée sur la figure 4.2. Nous pouvons tirer de cette mesure que la polarisation devient possible dès que l'on applique une tension de programmation de 12V. De plus, après $100\mu s$ et pour des tensions de programmation supérieures à 12V, la tension de seuil augmente presque linéairement avec le temps de programmation en échelle logarithmique, donc en réalité de façon exponentielle avec le temps. Nous remarquons également qu'à durée de programmation équivalente, une différence de

$1V$ de la tension de programmation se traduit par un décalage de près de $1V$ de la tension de seuil.

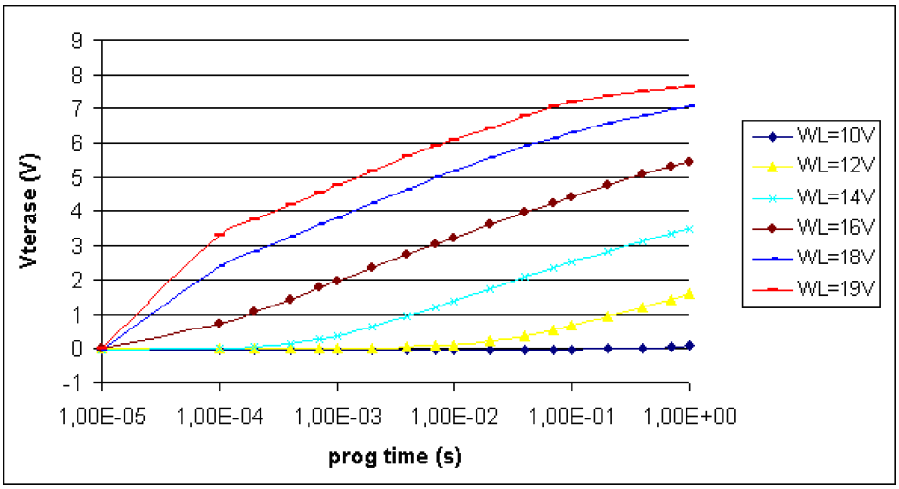


FIG. 4.2 – V_{Terase} en fonction du temps de programmation pour différentes valeurs de WL avant cyclage

La figure 4.3 représente la tension de seuil atteinte après $100\mu s$ de polarisation sur la grille de contrôle WL. La durée de programmation de $100\mu s$ correspond à l'objectif idéal de temps de programmation du produit.

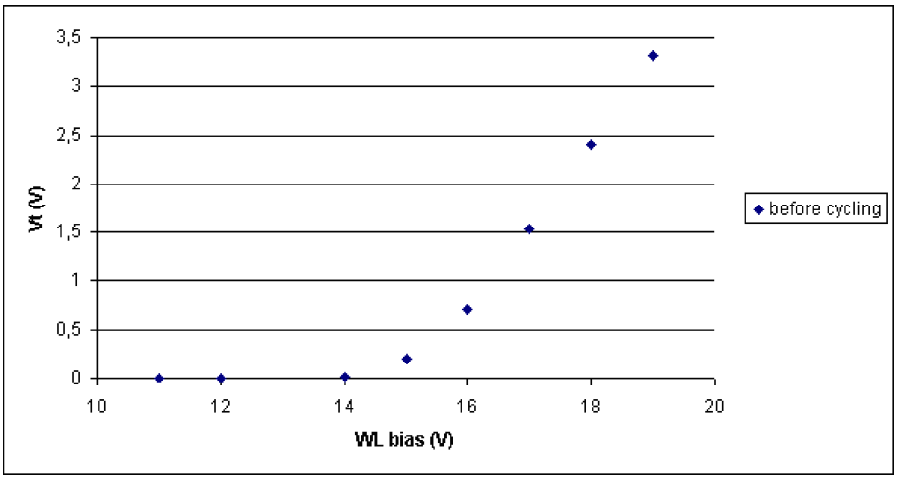


FIG. 4.3 – V_{Terase} après un stress de $100\mu s$ à différentes polarisations avant cyclage

Nous voyons que pour atteindre un V_T proche de $+4V$ avec un signal d'une durée de $100\mu s$, il faudrait appliquer une amplitude supérieure à $+20V$ sur la WL, ce qui n'est pas concevable en pratique du fait de la limitation des circuits périphériques de

génération des hautes tensions. La solution consiste donc à augmenter la durée de programmation à $1ms$ et à limiter la tension à $+18V$.

4.1.1.2 Perturbation avant cyclage

Nous rappelons que le program disturb est la perturbation causée sur une cellule à l'état effacé par une programmation sur une cellule partageant la même ligne de mot WL. Une tension de $2,5V$ sur le Drain de la cellule non-sélectionnée permet de limiter le champ aux bornes de l'oxyde tunnel sur cette cellule que l'on doit éviter de programmer, par opposition à la tension de 0V sur le Drain de la cellule en phase de programmation qui permet d'appliquer un fort champ électrique. Par couplage capacitif entre la grille de contrôle et la grille flottante, lorsque la grille de contrôle est polarisée à 18V, la grille flottante monte à $18 \times 0,6 = 10,8V$ donc avec la tension de $2,5V$ sur le drain de la cellule non-sélectionnée, la différence de potentiels aux bornes de l'oxyde tunnel est de $8,3V$, ce que nous avons cherché à reproduire avec une polarisation de 0V sur le drain. La différence de potentiel est alors identique pour une tension de $8,3V$ sur la grille flottante, soit $8,3/0,6 = 13,8V$ sur la grille de contrôle. Nous utiliserons donc des tensions de grille de contrôle de $+10V$ à $+14V$.

Pour évaluer ce "program disturb", les polarisations du tableau 4.1 reproduisant les conditions visant à empêcher une programmation non-souhaitée de la cellule ont été appliquées en utilisant des signaux carrés.

Contacts	**Tensions**
Drain	$0V$
Grille de Contrôle	de +10V à +14V
Source	flottante
Substrat	0V

TAB. 4.1 – Conditions de Perturbation

Idéalement, la cellule ne devrait subir aucune modification de sa tension de seuil au cours de la phase de programmation de la cellule partageant la même ligne de mot. Les mesures de décalage de la tension de seuil effacée sont reportées sur la figure 4.4.

Jusqu'à des tensions de 12V, le décalage du V_T est quasi négligeable, même pour des temps élevés (inférieur à $100mV$ pour $t = 30s$) comme résumé dans le tableau 4.2 . En revanche, à partir de $13V$, un véritable effet de la perturbation apparaît, qui s'accroît avec la tension de programmation. Nous nous intéressons particulièrement aux temps $100ms$ et $30s$ qui correspondent respectivement à la perturbation que reçoit une cellule après 100 phases de programmation de cellules voisines et à la perturbation d'un pire cas pour lequel la perturbation est très marquée.

Après $100ms$, le décalage de la tension est linéaire avec le logarithme du temps, donc varie en exponentielle du temps. En effet, si après $100ms$ pour une tension de

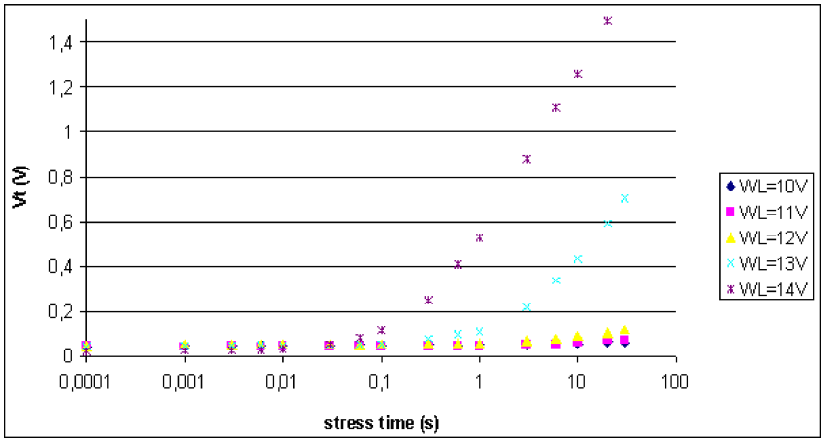


FIG. 4.4 – Mesure de la perturbation de la tension de seuil avant cyclage

Tension de WL	**Perturbation après** $100ms$	**Perturbation après** $30s$
10V	7mV	20mV
11V	0mV	27mV
12V	7mV	74mV
13V	8mV	660mV
14V	95mV	1,62V

TAB. 4.2 – Valeurs des Perturbations avant cyclage

$14V$ la perturbation induite n'est que d'environ $100mV$, après $30s$ elle est cette fois-ci de $1,62V$, ce qui ne serait pas acceptable pour un produit car conduirait à un basculement de l'état de la cellule. Le temps total de la perturbation pouvant être subie par une cellule doit donc être maintenu aux alentours de $100ms$ qui garantit un faible décalage du V_T et une conservation de l'état initial de la cellule perturbée.

4.1.2 Après cyclage

Le protocole de cyclage a été défini au chapitre 2.1. Un cyclage de 10.000 cycles a ensuite été réalisé conformément aux polarisations définies au paragraphe 4.1 avec les signaux carrés suivants :

- une tension de programmation $V_{prog} = 18V$ appliquée pendant une durée de 1ms,
- une tension d'effacement $V_{erase} = 16,5V$ appliquée pendant une durée de $80\mu s$

4.1.2.1 Détermination de l'efficacité de programmation après cyclage

Les mêmes mesures que dans le cas de la cellule non-cyclée ont été renouvelées avec des tensions de programmation V_{prog} variant de $+10V$ à $+19V$. Nous retrouvons quasiment les mêmes mesures de programmation en fonction du temps qu'avant le cyclage, comme le montre la figure 4.5. Nous pouvons également voir sur la figure 4.6 la tension de seuil atteinte après une phase de programmation d'une durée de $100\mu s$, elle aussi assez proche de celle obtenue avant cyclage.

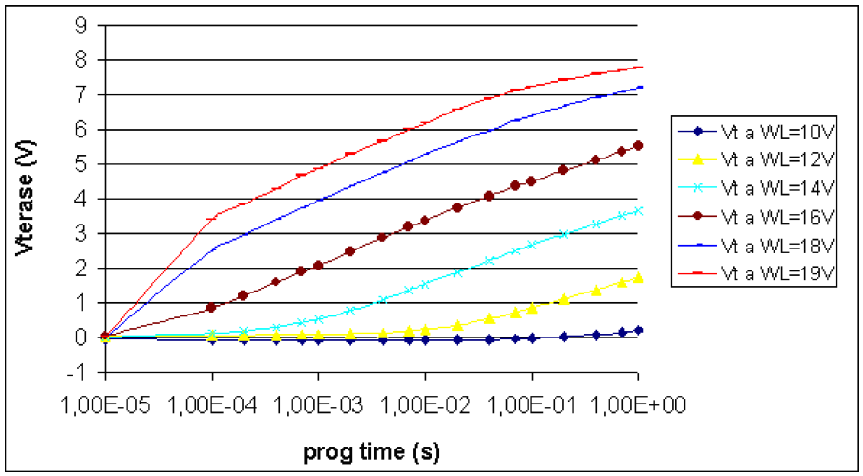


FIG. 4.5 – V_{Terase} en fonction du temps de programmation pour différentes valeurs de WL après cyclage

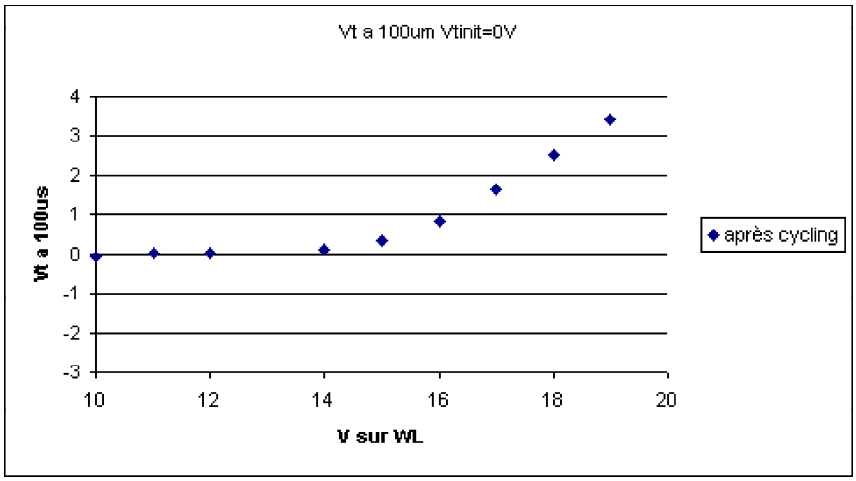


FIG. 4.6 – V_{Terase} après un stress de $100\mu s$ à différentes polarisations après cyclage

Pour comparer ces résultats avec ceux avant cyclage, nous avons rassemblé dans le tableau 4.3 la variation de la tension de seuil effacée après $1ms$ de programmation avant et après cyclage, soit la durée d'un signal de programmation. Nous pouvons remarquer que quelle que soit la tension appliquée sur la WL lors de la programmation, l'efficacité augmente après cyclage d'environ $100mV$ après $1ms$, ce qui est assez caractéristique de l'évolution de la tension de seuil programmée qui a souvent tendance à augmenter avec le nombre de cycles effectués. La programmation est donc plus efficace après cyclage, contrairement à l'effacement.

V_{prog}	ΔV_{Tprog} **après** $1ms$ **avant cyclage**	ΔV_{Tprog} **après** $1ms$ **après cyclage**
10V	0mV	0mV
12V	3mV	83mV
14V	376mV	534mV
16V	1,94V	2,08V
18V	3,82V	3,95V
19V	4,78V	4,88V

TAB. 4.3 – Comparaison des efficacités de programmation, avant et après cyclage.

Le cyclage n'a donc pas particulièrement modifié l'efficacité de programmation des cellules, avec une augmentation de $100mV$ de la tension de seuil après cyclage par rapport à celle atteinte avant cyclage, ce quel que soit le temps de programmation.

4.1.2.2 Perturbation après cyclage

Nous reproduisons après cyclage les mêmes mesures de program disturb que celles réalisées au paragraphe 4.1.1.2 avant cyclage. La figure 4.7 représentant ces dégradations montre une plus forte sensibilité aux perturbations après le cyclage.

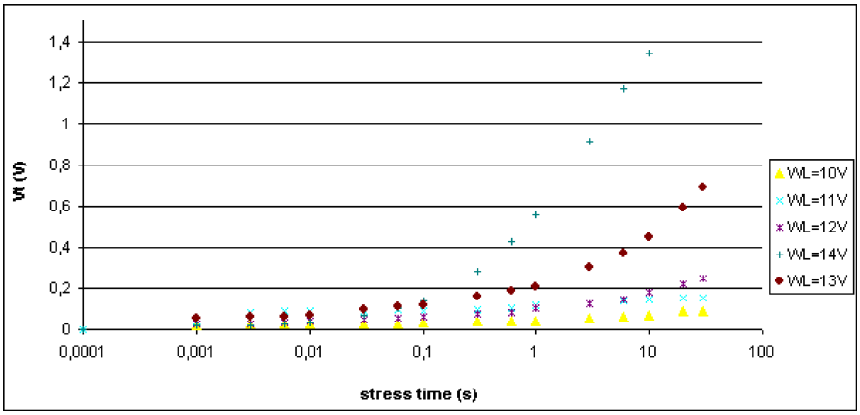


FIG. 4.7 – Mesure de la perturbation après cyclage

La comparaison des résultats est reportée dans le tableau 4.4 qui montre qu'il existe une importante dégradation de la résistance au disturb de grille après un cyclage de 10.000 cycles. Cette dégradation provoque des valeurs de disturb jusqu'à 12 fois plus grandes pour les tensions les plus élevées. Cela pourrait poser des problèmes dans la future technologie dont la durée de vie devra être d'au moins 100.000 cycles. Des modifications dans le process seront certainement nécessaires afin d'améliorer la tenue au disturb de grille.

Tension de WL	**Perturbation après** $100ms$	**Perturbation après** $30s$
10V	55mV	106mV
11V	57mV	143mV
12V	77mV	259mV
13V	134mV	706mV
14V	133mV	1,69V

TAB. 4.4 – Valeurs des perturbations après cyclage

Pour mieux évaluer l'influence de l'architecture des cellules, il est indispensable d'étendre ce type de mesures à d'autres technologies et notamment à des cellules Flash en architecture NAND.

4.2 Perturbations de grille sur cellule mémoire en architecture NAND S16

Nous utilisons dans cette étude de perturbation la cellule mémoire NAND à 16 cellules appelée "S16", que nous rappelons en figure 4.8.

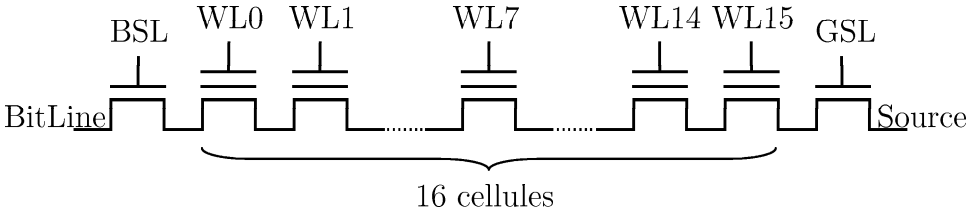


FIG. 4.8 – Structure mémoire S16

Nous réalisons sur cette cellule une mesure de program disturb, c'est-à-dire nous mesurons l'impact d'une programmation de la cellule sélectionnée sur une cellule inhibée à l'état effacé. Nous utilisons pour l'ensemble des mesures une cellule située au centre de la chaîne mémoire constituée de 16 cellules (WL0 à WL15), soit la cellule WL7. Habituellement, ces cellules sont celles qui en mesure présentent la meilleure

reproductibilité, contrairement aux cellules situées en extrémités de chaîne dans lesquelles peuvent intervenir des effets de bord.

4.2.1 Avant Cyclage

Si l'on applique un signal classique de programmation sur la cellule sélectionnée, la cellule inhibée reçoit également la forte polarisation positive sur sa grille de contrôle, leurs contacts de grille de contrôle étant communs, et peut subir une légère programmation, que nous pouvons observer sur la figure 4.9 par un décalage de la caractéristique I_{BL}-V_{WL} (courant de la ligne de bit en fonction de la polarisation de la ligne de mot).

A l'état vierge, les deux cellules ont exactement la même caractéristique. Il en est de même après l'effacement de chacune de ces cellules. Au cours de la phase de programmation avec un signal de programmation de 18V pendant $200\mu s$, la cellule sélectionnée passe d'une tension de seuil $V_{T_{erase}} = -4V$ à une tension de seuil $V_{T_{prog}} = +1,5V$ tandis que la cellule inhibée ayant la même ligne de mot qui est au départ également dans l'état effacé à $V_T = -4V$ subit un décalage de sa tension de seuil de $250mV$.

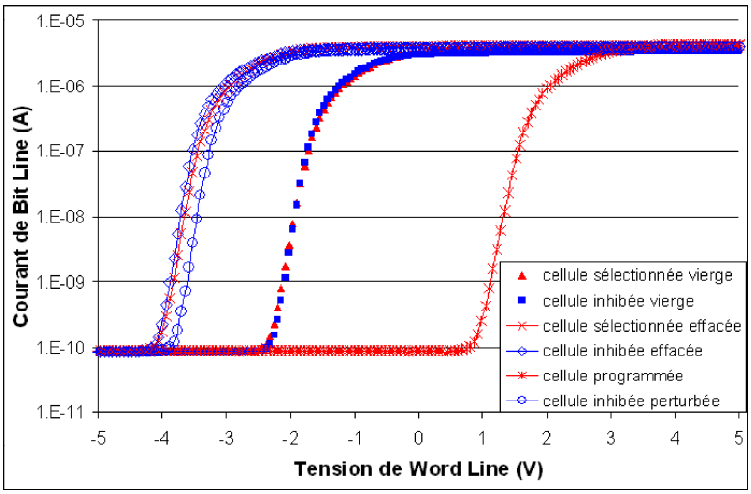


FIG. 4.9 – Mesure du disturb avant cyclage sur cellule S16

4.2.2 Après cyclage

Nous souhaitons reproduire cette mesure de perturbation de la cellule inhibée après l'avoir cyclée afin de vérifier la résistance aux perturbations d'une cellule avec un oxyde dégradé. Nous utilisons des signaux carrés de $200\mu s$ en programmation

et $1ms$ en effacement avec des polarisations dont les valeurs sont résumées dans le tableau 4.5.

Après avoir effectué 100.000 cycles sur la cellule inhibée selon ces conditions, nous mesurons à nouveau le niveau de perturbation de la cellule inhibée en vue de comparer les valeurs avec celles obtenues avant le cyclage, afin de mettre en évidence l'effet du cyclage.

Contact	Tensions de Programmation	Tensions d'Effacement
Bit Line	$0V$	$V_{erase} = 17V$
Word Line 7	$V_{prog} = 17V$	$0V$
Autres Word Lines	$V_{pass} = 8V$	$0V$
Source	$4,5V$	$V_{erase} = 17V$
Substrat	$0V$	$V_{erase} = 17V$
Transistor BSL	$1,5V$	$V_{erase} = 17V$
Transistor GSL	$0V$	$V_{erase} = 17V$

TAB. 4.5 – Valeurs des Tensions utilisées pour le cyclage

La figure 4.10 présente la comparaison avant et après cyclage des niveaux de perturbation sur 6 sites différents de la zone centrale du wafer.

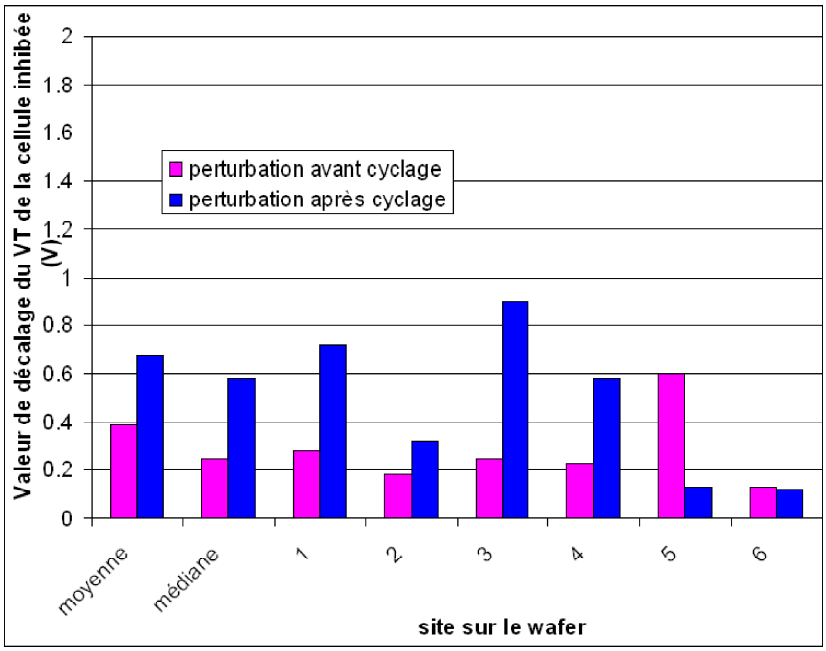


FIG. 4.10 – Comparaison du disturb avant et après cyclage sur cellule S16

En règle générale, le niveau de perturbations après cyclage est environ deux fois plus élevé qu'avant cyclage, excepté un des sites mesurés pour lequel la dégradation est moins élevée après qu'avant cyclage et un second pour lequel la dégradation est identique. Concernant le site n°5, la cellule présente un courant de saturation après cyclage très inférieur à celui avant cyclage ce qui est caractéristique d'une cellule très dégradée, proche du non-fonctionnement, ce qui peut expliquer un fonctionnement atypique. Concernant le site n°6 avec le même niveau de perturbation avant et après cyclage, aucune explication n'a pu être avancée.

En plus des mesures de tenue à la perturbation de la cellule inhibée, nous avons cherché à étudier la tenue de cette cellule au cyclage de la cellule sélectionnée, sachant que la cellule inhibée subit des successions d'effacement classique et de légères programmation comme nous venons de le mettre en évidence.

4.3 Problématique de la dégradation des cellules inhibées

4.3.1 Dégradations observées

Les conditions de programmation au cours de cette étude sont les mêmes que celles utilisées lors des cyclages entre les mesures de perturbations de la cellule inhibée, décrites dans le tableau 4.5. La seule différence entre les cellules inhibée et sélectionnée est la tension appliquée sur la BL lors de la programmation : $0V$ pour une cellule sélectionnée et $V_{inh} \approx 3V$ pour une cellule inhibée. Cette polarisation V_{inh} induit une hausse de la tension dans le canal de la chaîne inhibée, phénomène appelé "boostage du canal", afin de diminuer le champ électrique aux bornes de l'oxyde tunnel et d'empêcher la programmation [Suh'95][Satoh'99]. 100.000 cycles d'écriture/effacement ont été faits sur des cellules avec les conditions suivantes : $200\mu s$ à $17V$, sur la WL en programmation ou sur le substrat en effacement. Cependant, pour certains splits de process, il arrive que lorsque l'on observe sur des cellules sélectionnée et inhibée la variation de la mesure médiane (sur 7 composants) de la tension de seuil V_T pendant le cyclage, la cellule sélectionnée montre la fermeture habituelle de la fenêtre de programmation, comme cela est le cas sur la figure 4.3.1, tandis que la cellule inhibée, qui ne devrait montrer aucune variation du V_T, souffre d'un décalage de $500mV$ de la tension de seuil du niveau effacé.

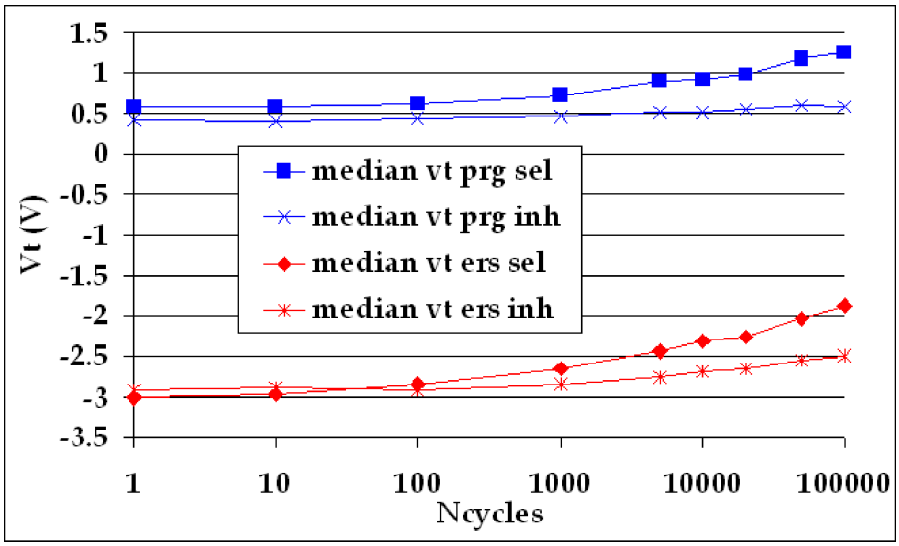


FIG. 4.11 – Un exemple de dégradation des cellules sélectionnées et inhibées au cours du cyclage, avant optimisation du process.

4.3.2 Hypothèses de mécanismes de dégradation de la cellule inhibée

Afin d'expliquer ce phénomène de dégradation des cellules inhibées lors du cyclage de la cellule sélectionnée, plusieurs mécanismes peuvent être supposés comme étant la cause de cette dégradation :

- une injection de porteurs chauds à travers l'oxyde lors de la programmation de la cellule sélectionnée comme cela a pu être reporté par Lee [Lee'06]
- une injection de charges par effet tunnel de type Fowler-Nordheim à bas champs lors de la phase d'inhibition de la cellule
- d'autres mécanismes qui resteraient à identifier

La simulation TCAD bidimensionnelle semble être une bonne voie pour étudier ce phénomène qui pose des problèmes évidents de fiabilité, c'est pourquoi nous allons maintenant détailler la mise en œuvre d'une simulation de type TCAD la plus réaliste possible afin de reproduire tout le fonctionnement de la cellule étudiée.

4.4 Simulation bidimensionnelle d'une chaîne NAND à 1 bit

Nous pouvons reproduire le processus complet de fabrication des cellules qui ont été utilisées précédemment pour les mesures de dégradation sur une chaîne NAND

à 1 bit. Nous pourrons ainsi simuler électriquement les conditions de programmation et d'effacement, aussi bien sur cellule sélectionnée que sur cellule inhibée. Nous pourrons ainsi observer et identifier dans la suite de cette étude le(s) mécanisme(s) responsable(s) des dégradations mesurées.

Nous allons tout d'abord présenter les différentes étapes successives nécessaires à l'obtention de la structure de simulation.

4.4.1 Simulation Process

Le premier travail consiste à reproduire en simulation toutes les étapes process effectivement réalisées sur silicium. Nous disposons pour cela de l'outil "Sentaurus Process". Ce logiciel permet de simuler l'ensemble des actions intervenant dans les méthodes de fabrication des circuits intégrés, telles que les définitions de masque, les dépôts, les photolithographies, les gravures, les implants de dopants, les diffusions, ...Chacune de ces étapes est rigoureusement définie par un jeu complet de paramètres.

A titre d'exemple, pour réaliser une implantation de dopants, nous devons renseigner :
- l'espèce à implanter (Bore, Arsenic, Phosphore, ...)
- le masque à travers lequel se fait l'implantation
- la dose de dopants (en cm^{-2})
- l'énergie d'implantation (en keV)
- l'angle du faisceau d'implantation par rapport à la normale du wafer ou *tilt* (en °)
- la rotation de ce faisceau dans le plan du wafer (en °)

ou encore pour une diffusion :

- le(s) type(s) de gaz sous lequel a lieu la diffusion
- le(s) flux de ce(s) gaz
- la pression de ce(s) flux de gaz
- la cinétique de montée en température
- la durée de diffusion à température maximum
- la cinétique de baisse en température

Le principal travail au cours de la mise en place de cette simulation est de suivre les différentes étapes de fabrication, fournies par l'entreprise, et de mettre à jour l'ensemble des paramètres avec les conditions réelles du wafer sur lequel ont été réalisées les mesures.

Nous avons ainsi reproduit chacune des multiples étapes de fabrication des structures qui ont été mesurées sur silicium.

La figure 4.12 montre la structure que nous obtenons en sortie du logiciel "Sentaurus Process" après un temps de simulation d'environ deux jours.

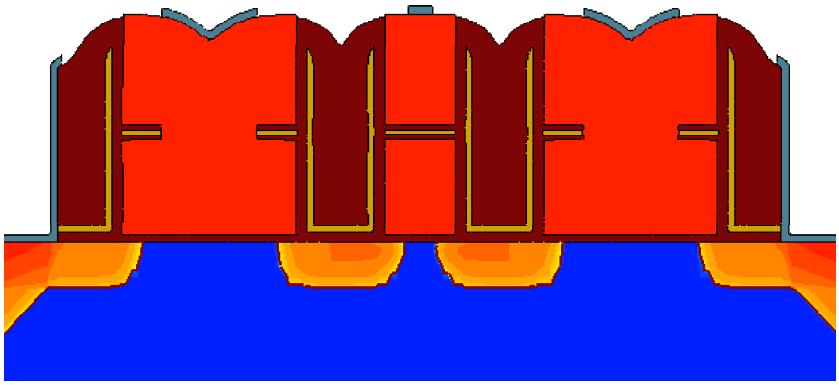


FIG. 4.12 – Structure obtenue en sortie de la simulation process.

Nous devons néanmoins mailler de façon pertinente cette structure et calibrer les grandeurs électriques en vue des simulations électriques à suivre.

4.4.2 Simulation de la structure

Cette partie, réalisée via l'outil "Sentaurus Structure Editor" nous a permis de définir les contacts et de redéfinir le maillage, notamment en diminuant la dimension des mailles à l'intérieur du canal de la cellule mémoire et dans les zones d'implants "Source" et "Drain". Nous pourrons grâce à cela visualiser tout mécanisme éventuel qui apparaîtrait dans ces zones les plus probables de localisation du phénomène. La structure finale avec le maillage "optimisé" permettant une convergence des simulations dans un temps de calcul raisonnable est reproduite en figure 4.13.

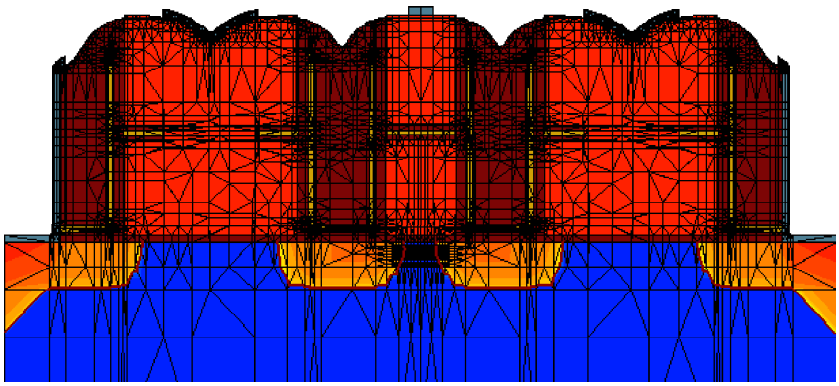


FIG. 4.13 – Structure finale obtenue.

4.4.3 Simulation électrique

Avant d'identifier un quelconque phénomène, nous devons encore calibrer électriquement notre cellule mémoire, en vérifiant les caractéristiques du courant de ligne de bit I_{BL} en fonction de la tension appliquée sur la ligne de mot V_{WL} d'un point mémoire à la tension de seuil naturelle (pour $Q_{FG} = 0$), à la tension de seuil programmée ($V_{T_{prog}} \approx 1,5V$, soit une charge dans la grille flottante $Q_{FG} = -3,5e16$ C) et à la tension de seuil effacée ($V_{T_{erase}} \approx -3V$, soit une charge dans la grille flottante $Q_{FG} = 9e16$ C). Il est aussi possible de calibrer la tension de seuil des transistors de sélection qui ont une influence sur le courant dans la chaîne totale. La figure 4.14 reproduit l'ensemble de ces caractéristiques, mesurées et simulées, montrant un très bon accord entre la simulation et les mesures sur silicium.

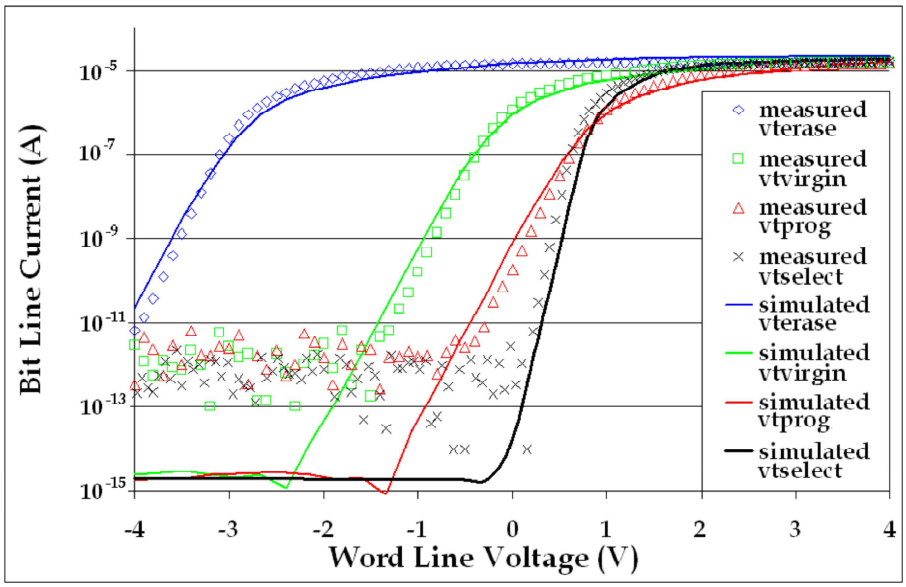


FIG. 4.14 – Validation des paramètres process par simulation électrique des caractéristiques I_{BL} en fonction de V_{WL} naturelle, programmée et effacée de la cellule mémoire ainsi que du transistor de sélection.

Nous devons aussi calibrer les paramètres d'injection de courant à travers l'oxyde tunnel lors de la phase de programmation. Nous avons pour cela extrait à l'aide d'un programme Mathcad les paramètres caractéristiques de l'équation Fowler-Nordheim, A et B, à partir de mesures de courant sur des capacités de grande superficie, selon la méthode décrite au paragraphe 3.3.3 page 89. Les courbes de courant de programmation sont données dans la figure 4.15.

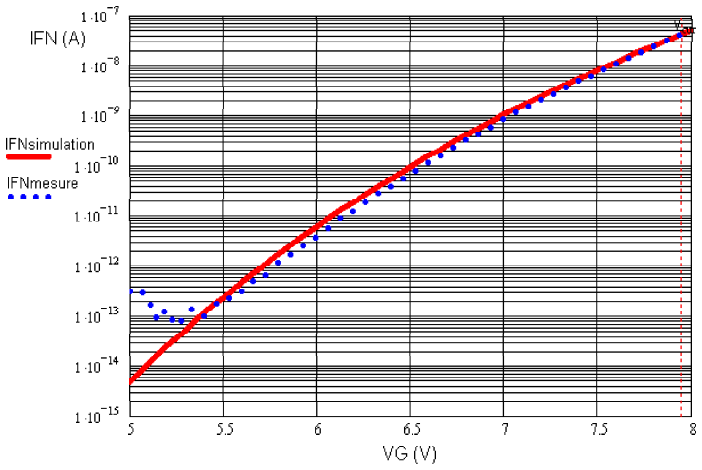


FIG. 4.15 – Courbes mesurée et simulée du courant Fowler-Nordheim en programmation.

Connaissant nos paramètres d'injection Fowler-Nordheim en programmation, nous pouvons désormais les entrer dans nos simulations électriques pour reproduire des courbes de programmation en fonction du temps (cf. figure 4.16).

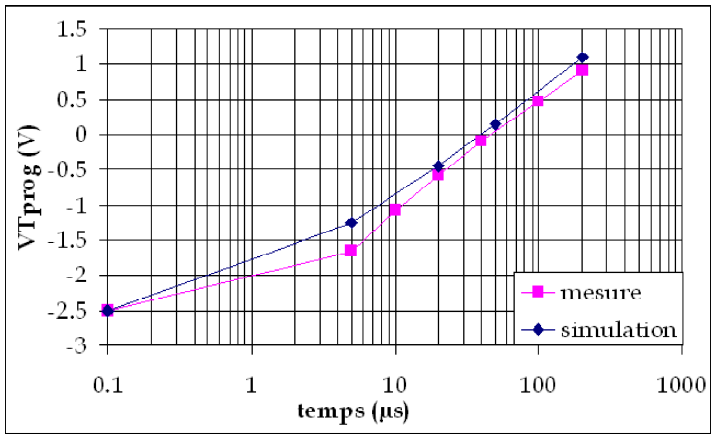


FIG. 4.16 – Validation des paramètres par simulation électrique de l'efficacité d'injection en programmation.

Toutes ces simulations électriques concordent très bien avec les mesures sur Silicium. Notre structure simulée peut donc être considérée comme étant calibrée et peut être utilisée pour des simulations électriques des phases de programmation aussi bien sur la cellule sélectionnée que sur la cellule inhibée. Cependant, afin d'encore affiner notre simulation nous allons chercher à prendre en compte les capacités de couplage qui existent entre les cellules voisines à l'intérieur d'une matrice mémoire.

4.5 Capacités de couplage entre cellules mémoires Flash en architecture NAND

4.5.1 Définition des capacités de couplage

Les capacités "parasites" de couplage sont les capacités qui existent entre les cellules à l'intérieur de la matrice mémoire et qui conduisent lors de l'utilisation de celles-ci à des modifications des polarisations appliquées et par conséquent à une perturbation des tensions de seuil des cellules, voire à des injections parasites de charges.

4.5.2 Bibliographie sur les capacités de couplage

De nombreuses études ont été menées sur le développement de simulations en trois dimensions afin de prédire les valeurs des perturbations sur la tension de seuil induites par ces capacités "parasites"[Lee'02][Lee'04][Ghetti'05]. Du fait de l'impossibilité physique de mesurer directement les valeurs de ces capacités, la méthode la plus utilisée repose sur des simulations TCAD de matrices de cellules. Dans tous les cas, une matrice 2x2 ou 3x3 de cellules est mise en place puis les capacités de couplage entre grilles flottantes (C_{FGx} en x, C_{FGy} en y et C_{FGxy} en xy), entre grille flottante et grille de contrôle voisine (C_{FGCG} en y) ainsi qu'entre grilles de contrôle voisines (C_{CGCG} en y) sont extraites. La figure 4.17 représente une matrice 3x3 ainsi que toutes ces capacités de couplage. La capacité d'ONO C_{ONO} entre la grille flottante et la grille de contrôle appartenant à une même cellule et la capacité tunnel C_{Tun} correspondant à l'oxyde tunnel, peuvent également être extraites à partir de cette simulation.

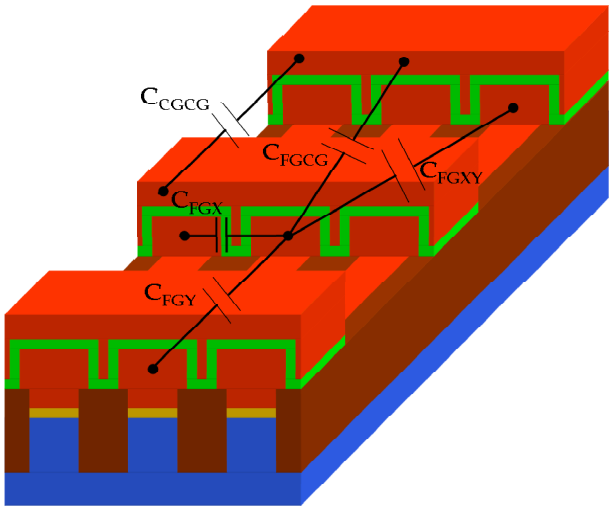


FIG. 4.17 – Schéma des capacités parasites dans une matrice 3x3.

Ces extractions de capacités "parasites" ont toujours pour but d'évaluer la perturbation de la tension de seuil due au couplage. En effet, dans la vision conventionnelle, la tension de grille flottante V_{FG} est reliée à celle de la grille de contrôle V_{CG} par l'expression :

$$V_{FG} = \gamma . V_{CG} \tag{4.1}$$

où γ est le coefficient de couplage entre la grille de contrôle et la grille flottante, défini par :

$$\gamma = \frac{C_{ONO}}{C_{TOT}} \tag{4.2}$$

où C_{ONO} représente la capacité entre la grille de contrôle et la grille flottante, et C_{TOT} est la capacité totale de la grille flottante exprimée par :
$C_{TOT} = C_{ONO} + C_{Tun}$.

Cependant, du fait de la réduction continue des dimensions des cellules, l'impact des cellules voisines sur la cellule étudiée semble devoir jouer un rôle et devient non-négligeable. Ainsi, si l'on considère la grille flottante de la cellule au centre de la matrice 3x3 de la figure 4.17, elle est couplée capacitivement dans la direction x avec une cellule à droite et une cellule à gauche, dans la direction y avec une cellule devant, une cellule derrière et les deux lignes de grille de contrôle et dans les directions xy une cellule dans chacun des 4 angles.
Le coefficient de couplage pour cette grille flottante devient donc :

$$\gamma_{FG_i} = \frac{C_{FG_i}}{C_{ONO} + C_{Tun} + 2C_{FGx} + 2C_{FGy} + 4C_{FGxy} + 2C_{FGCG}} \tag{4.3}$$

soit encore en terme de changement de polarisation due à un changement de ΔV_T des cellules voisines :

$$\Delta V_{FG} = \sum_i \frac{C_{FG_i}}{C_{ONO} + C_{Tun} + 2C_{FGx} + 2C_{FGy} + 4C_{FGxy} + 2C_{FGCG}} . \Delta V_T \tag{4.4}$$

où i représente les directions x, y ou xy.
Nous voyons donc que toute modification de la tension de seuil d'une cellule induit sur les cellules voisines une modification proportionnelle de la tension de seuil. Cela pose des problèmes évidents de fiabilité qui doivent être minimisés autant que possible.

Nous allons également étudier ces capacités "parasites" dans le cadre de la cellule développée par l'entreprise Atmel. Il s'agit d'une cellule mémoire Flash en architecture NAND, appartenant au nœud technologique $90nm$, c'est-à-dire que la largeur de la grille est de $90nm$. L'épaisseur de l'oxyde tunnel est de $8nm$ et celle de l'ONO est de $20nm$.

4.5.3 Simulation tri-dimensionnelle d'une matrice 3x3 de cellules mémoires

Nous avons pour cette étude mis en place une simulation tri-dimensionnelle d'une matrice 3x3 de nos cellules mémoires avec des implants de Source et Drain simplifiés pour limiter le temps de calcul, sachant que ces implants n'ont qu'un impact très limité sur les couplages entre grilles de contrôle et grilles flottantes. Nous avons donc fixé des implants constants. Pour encore davantage limiter le temps de calcul, la matrice complète est obtenue par réflexions successives d'un quart de cellule. La matrice 3x3 représentée en figure 4.18, réalisée grâce à l'outil de simulation TCAD "Sentaurus Structure Editor", et dont une vue sélective du polysilicium est visible en figure 4.19, montre bien la présence de 9 cellules élémentaires.

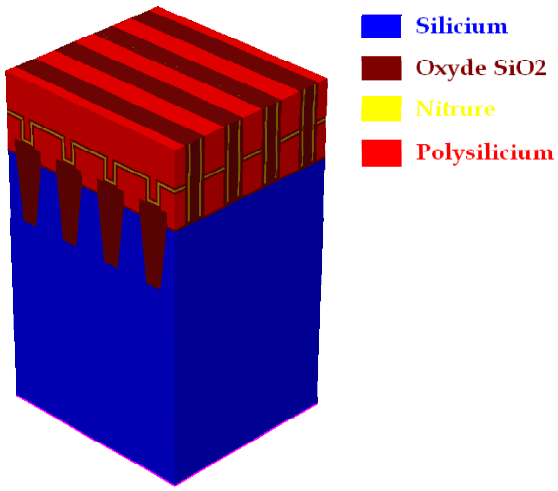


FIG. 4.18 – Matrice 3x3 de cellules mémoires Flash en architecture NAND, simulée en trois dimensions.

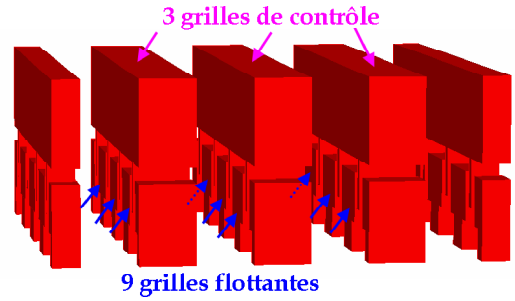


FIG. 4.19 – Matrice 3x3 de cellules mémoires Flash en architecture NAND, vue sélective du Polysilicium.

La simulation de la matrice 3x3 nous a permis d'extraire les valeurs de capacités, rassemblées dans le tableau 4.6, en appliquant une petite modulation d'amplitude $1\mu V$ à la fréquence de $10kHz$ sur la grille flottante au centre de la matrice et en appliquant $0V$ sur les cellules voisines. Ces valeurs sont en accord avec les ordres de grandeurs des capacités de couplage "parasites" que l'on peut trouver généralement dans la bibliographie pour des technologies 130nm [Lee'02].

Nous reviendrons sur la cohérence de ces valeurs un peu plus tard dans cette étude.

Capacité	**Valeur (aF)**
C_{FGx}	2.394
C_{FGy}	5.065
C_{FGxy}	0.391
C_{FGCG}	2.199
C_{CGCG}	5.401
C_{ONO}	63.462
C_{Tun}	48.455

TAB. 4.6 – Valeurs des capacités extraites dans Sentaurus.

Si l'on se réfère à la formule (4.4), ces valeurs de capacités conduisent aux décalages de V_{FG} sur cellule voisines du tableau 4.7, pour un décalage de la tension de seuil $\Delta V_T = 4,5V$ sur la cellule sélectionnée.

Capacité parasite	ΔV_{FG} **(mV) induit**
C_{FGx}	81
C_{FGy}	172
C_{FGxy}	13
C_{FGCG}	75

TAB. 4.7 – Valeurs des modifications de tensions de seuil dues aux capacités extraites dans Sentaurus.

Nous voyons donc que si l'on programme chacune des cellules voisines d'une même cellule, placée au centre d'une matrice 3x3, le décalage de la tension de la grille flottante de cette cellule peut atteindre $2\times81+2\times172+4\times13+6\times75 = 1008mV \approx 1V$.

4.5.4 Méthode de mesure indirecte des capacités de couplage.

Afin de valider les valeurs de capacités de couplage "parasites" extraites à partir du logiciel de simulation TCAD, une mesure était indispensable.

La seule mesure possible à partir des structures de test développées a été une mesure indirecte basée sur plusieurs structures de même type mais qui diffèrent entre elles d'une seule des capacités de couplage mises en jeu dans la matrice mémoire.

4.5.4.1 Structures de test utilisées dans la mesure indirecte des capacités de couplage

Trois structures de test, appelées A, B et C dans la suite de ce manuscrit, ont permis la mesure indirecte des capacités de couplage et seront détaillées tour à tour. Ces structures sont constituées de 352 lignes de polysilicium imbriquées, chacune d'une longueur de 740μm, même si nous les représenterons comme deux cellules placées côte à côte, nous permettant d'avoir une grande surface et donc des capacités suffisamment grandes pour être mesurées.

Structure de test A

La structure de test A, représentée en figure 4.20, peut être vue comme une configuration classique de deux cellules identiques placées côte à côte. En polarisant la zone active à la masse GND, les seules capacités restantes dans la structure sont les capacités $C1$ représentant le couplage entre les deux grilles de contrôle, $C2 = C4$ les capacités d'ONO et $C3$ la capacité entre les grilles flottantes dans la direction y.

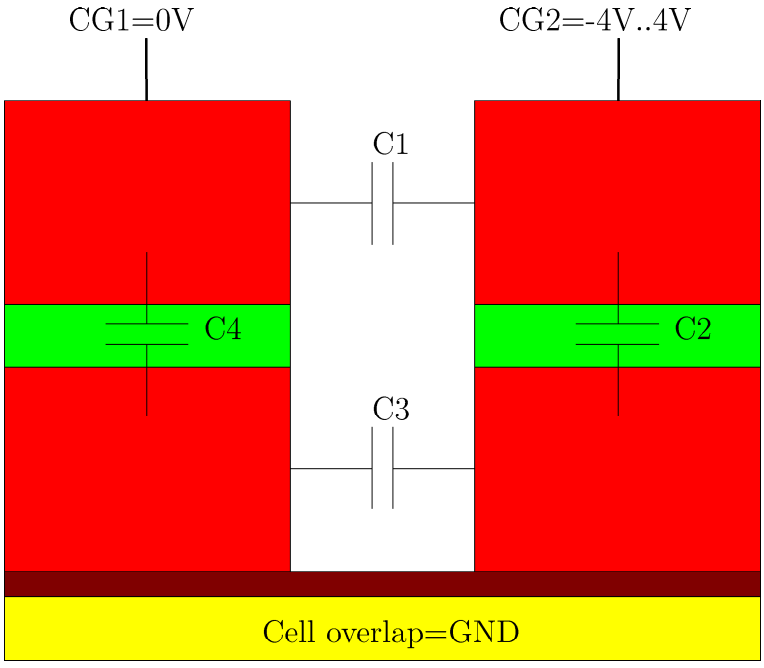


FIG. 4.20 – Structure de test A.

La capacité équivalente $C_{mesure1}$, mesurée en faisant varier la tension sur la grille de contrôle de la seconde cellule de $-4V$ à $4V$ vaut :

$$C_{mesure1} = C_1 + \frac{C2.C3.C4}{C2.C3 + C2.C4 + C3.C4}$$

Soit comme $C2 = C4 = C_{ONO}$:

$$C_{mesure1} = C_1 + \frac{C4.C3}{2.C3 + C4} = C_{CGCG} + \frac{C_{ONO}.C_{FGy}}{2.C_{FGy} + C_{ONO}} \tag{4.5}$$

Structure de test B

La structure B, représentée en figure 4.21, est similaire à la structure A présentée précédemment mis à part le fait que la grille de contrôle et la grille flottante de la cellule 2 sont court-circuitées. La capacité d'ONO $C2$ n'existe donc plus dans la structure, seules restent les capacités $C1$, $C3$ et $C4$.

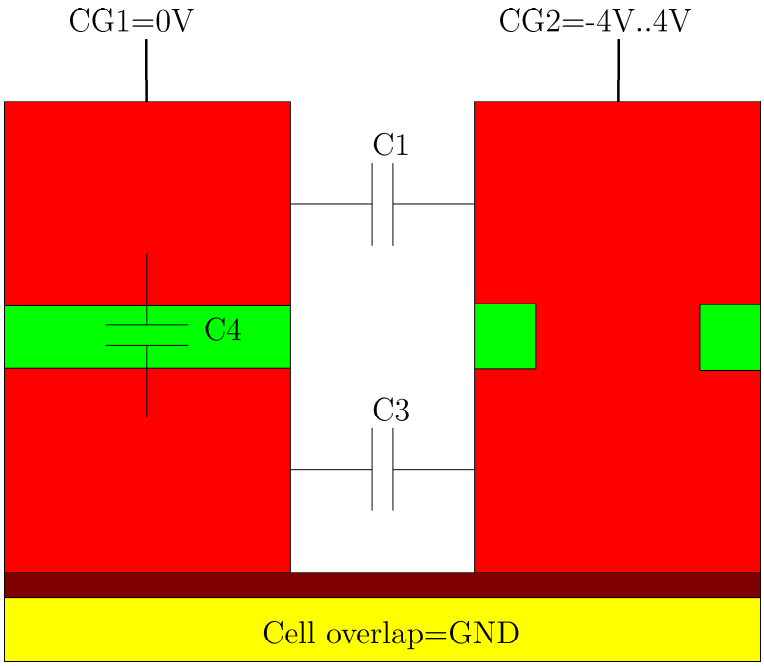


FIG. 4.21 – Structure de test B.

La capacité équivalente $C_{mesure2}$ vaut :

$$C_{mesure2} = C1 + \frac{C3.C4}{C3 + C4} = C_{CGCG} + \frac{C_{ONO}.C_{FGy}}{C_{FGy} + C_{ONO}} \tag{4.6}$$

Structure de test C

La structure C, représentée en figure 4.22, consiste en 2 cellules mises côte à côte et où les grilles de contrôle sont directement reliées avec les grilles flottantes, cela sur les deux cellules. Le circuit équivalent est donc celui de deux capacités mises en parallèles.

La capacité équivalente $C_{mesure3}$ vaut :

$$C_{mesure3} = C1 + C3 = C_{CGCG} + C_{FGy} \tag{4.7}$$

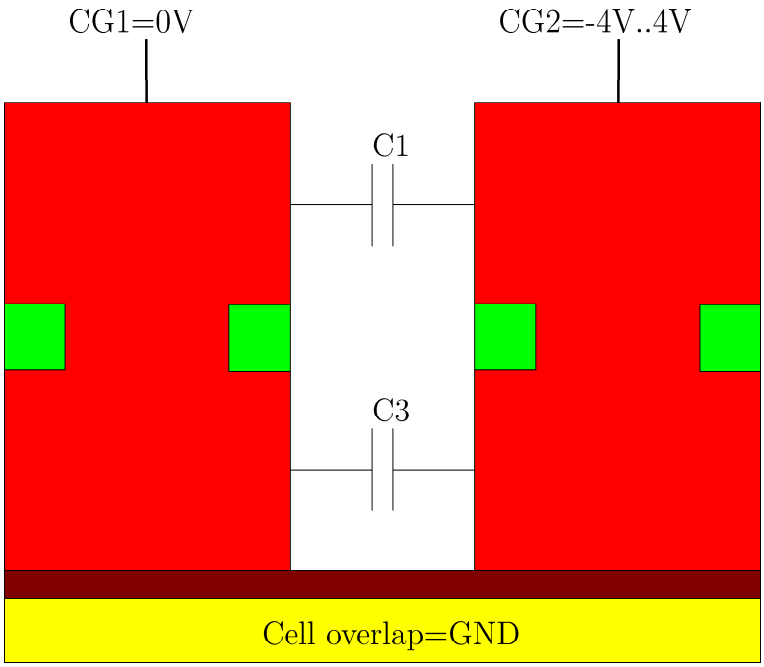


FIG. 4.22 – Structure de test C.

4.5.4.2 Mesures réalisées sur les structures de test A, B et C.

Des mesures, visibles en figure 4.23, ont été réalisées sur chacune des trois structures de test A, B et C.

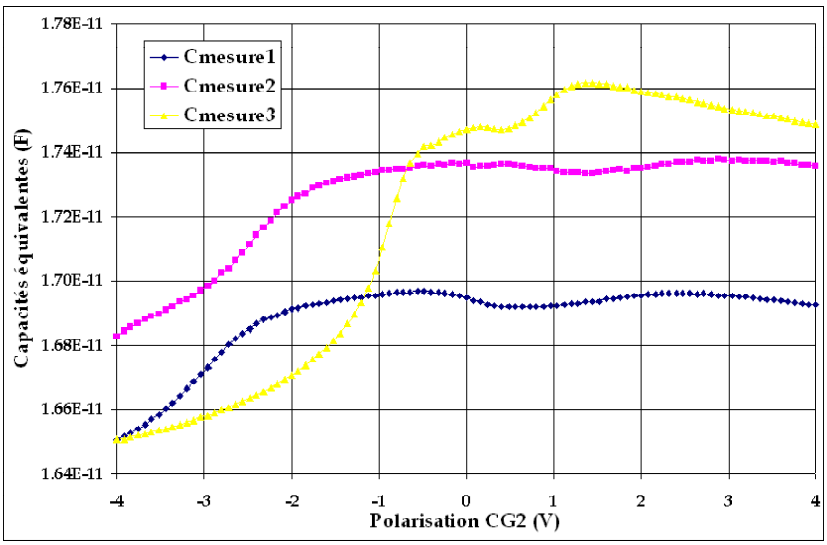


FIG. 4.23 – Courbes Cmesure1 (structureA), Cmesure2 (structureB) et Cmesure3 (structureC).

A partir de ces mesures, il est possible de prendre la valeur à $0V$ qui correspond à la polarisation $1\mu V$ utilisée en simulation[1] pour chacune des capacités mesurées, dont les valeurs sont reportées dans le tableau 4.8.

Capacité	**Valeur moyenne (aF/bit)**
$C_{mesure1}$	$8,47$
$C_{mesure2}$	$8,65$
$C_{mesure3}$	$8,75$

TAB. 4.8 – Valeurs des capacités équivalentes mesurées.

En reprenant l'expression de $C_{mesure3}$ donnée en (4.7), si l'on connaît le rapport entre les deux capacités (donc le rapport entre les deux surfaces respectives) C_{CGCG} et C_{FGy}, il est possible d'extraire directement la valeur de ces capacités.

Un calcul des surfaces (en prenant la largeur de la grille de contrôle égale à celle de la grille flottante W_{poly}) , à partir des dimensions données sur la figure 4.24, fournit la relation :

$$C_{CGCG} = 0.524 \times (C_{FGy} + C_{CGCG}) = 0.524 \times C_{mesure3} \tag{4.8}$$

d'où $C_{CGCG} = 4,58aF/bit$ et $C_{FGy} = 4,16aF/bit$

[1] ces valeurs seront données en aF/bit, c'est-à-dire que la capacité totale mesurée est ramenée à une surface unitaire puis multipliée par la surface latérale d'une cellule élémentaire. Pour rappel, $1aF = 10^{-18}F$.

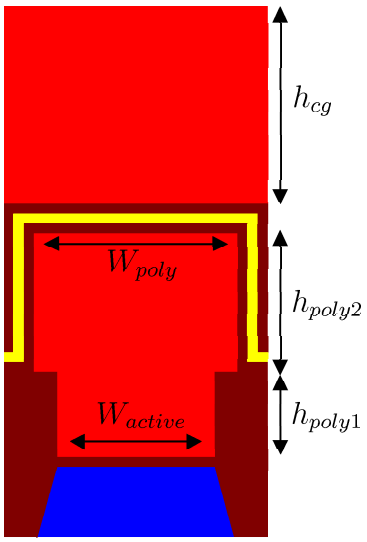


FIG. 4.24 – Vue en simulation TCAD des surfaces des capacités inter-grilles-flottantes et inter-grilles-de-contrôle.

4.5.5 Comparaison des valeurs simulées, mesurées, calculées et publiées

En plus des valeurs de ces capacités parasites obtenues par simulation TCAD et par mesures, nous pouvons calculer les valeurs de certaines de ces capacités à partir d'un calcul géométrique "simple" en utilisant l'équation $C_{ox} = \frac{\epsilon_0.\epsilon_{ox}.S_{ox}}{t_{ox}}$. Nous pouvons aussi comparer dans le tableau 4.9 les valeurs obtenues à celles que l'on peut trouver dans la littérature [Lee'02], en gardant à l'esprit que ce sont uniquement les ordres de grandeurs qu'il faut alors comparer du fait des différences de technologies, de géométries, ...

Capacité	Mesures	Simulation 3D	Calcul	Littérature 130nm
C_{CGY}	4,514	5,401	5,985	-
C_{FGY}	4,234	5,065	5,429	7,4
C_{FGX}	-	2,394	4,790	1,7
C_{FGCG}	-	2,199	-	5,2
C_{FGXY}	-	0,391	-	0,013
C_{ONO}	118,102	63,460	82,135	120
C_{TUN}	-	48,455	29,136	53

TAB. 4.9 – Comparaison des valeurs des capacités parasites, en aF/bit.

L'ensemble des valeurs extraites sur notre technologie sont en bon accord entre

elles (mesure, simulation et calcul direct). Le fait que la capacité C_{ONO} soit plus importante en mesure est due à une géométrie un peu particulière de la structure de test avec des grandes lignes qui ne sont pas similaires à une suite de composants isolés. Nous pouvons également remarquer que les ordres de grandeurs sont tout à fait cohérents avec ce que l'on peut trouver dans la littérature pour une technologie $130nm$ dans laquelle les valeurs ont tendance à être légèrement plus élevées du fait de la plus grande superficie des cellules par rapport à une technologie $90nm$.

4.6 Prise en compte des effets des capacités parasites

4.6.1 Prise en compte dans la simulation bi-dimensionnelle

Nous pouvons remarquer sur la figure 4.25 la présence de quatre contacts additionnels qui nous permettent d'introduire dans le simulateur électrique les quatre capacités de couplage entre cellules voisines dans les directions X et Y dans la matrice, extraites dans le paragraphe 4.5.5.

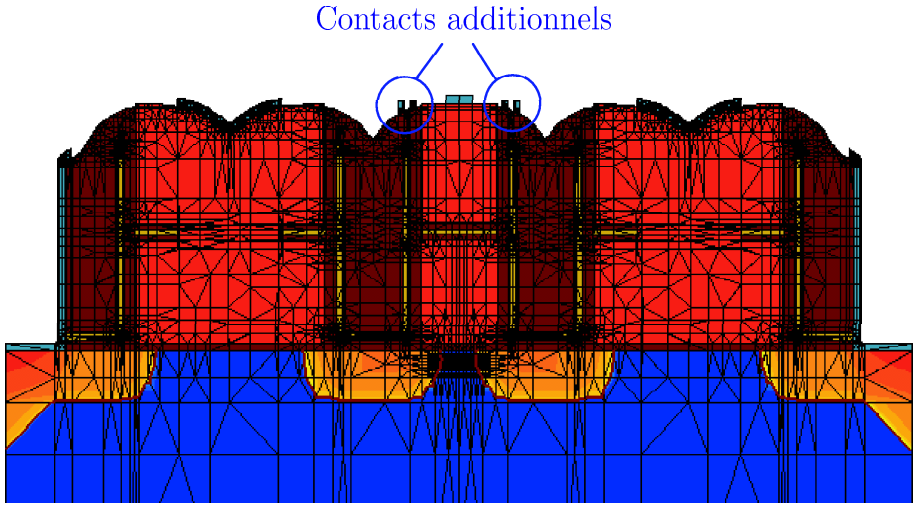


FIG. 4.25 – Structure finale avec les contacts additionnels.

4.6.2 Influence des capacités parasites

Ces capacités parasites, extraites sur une simulation en trois dimensions, ont donc pu être injectées dans une simulation en deux dimensions. Nous avons ainsi une prise en compte d'effets tridimensionnels tout en gardant la simplicité et la rapidité de simulations bidimensionnelles. Nous avons réalisé des simulations sans capacité parasite, puis en prenant en compte les capacités dans la direction X, puis dans la

direction Y, puis dans les deux directions X et Y. Les charges obtenues sont présentées sur la figure 4.26 en fonction du temps de programmation. Nous pouvons alors vérifier sur la figure 4.26 l'impact de ces capacités parasites sur la programmation de la cellule sélectionnée.

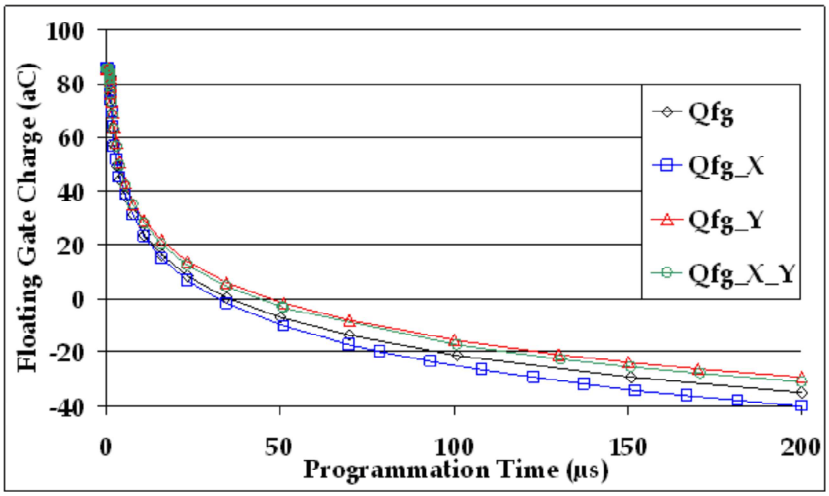


FIG. 4.26 – Evolution de la charge dans la grille flottante en fonction du temps de programmation, en prenant en compte : aucune capacité parasite (Qfg), les capacités parasites dans la direction X (Qfg_X), dans la direction Y (Qfg_Y) ou dans les deux directions X et Y (Qfg_X_Y).

Après la durée d'un signal de programmation classique, soit $200\mu s$, la charge stockée dans la grille flottante sans prendre en compte de capacité parasite est $Q_{FG} = -35,1aC$, ce qui équivaut à une tension de seuil programmée $V_{T_{prog}} = 1,6V$, alors que si l'on prend en compte les capacités parasites dans les directions X et Y, on obtient une charge finale $Q_{FG} = -31,2aC$, soit une tension de seuil programmée $V_{T_{prog}} = 1,45V$. Nous voyons donc que la prise en compte de ces capacités de couplage parasite introduit une diminution de $150mV$ du niveau programmé. Il semble donc judicieux de réaliser l'ensemble de nos simulations électriques ultérieures avec la prise en compte de ces capacités pour être au plus prêt de la réalité.

4.7 Identification du mécanisme de dégradation des cellules inhibées

On rappelle sur la figure 4.27 que la dégradation de la tension de seuil qui peut intervenir dans des conditions process non-optimisées sur les cellules inhibées est d'environ $500mV$.

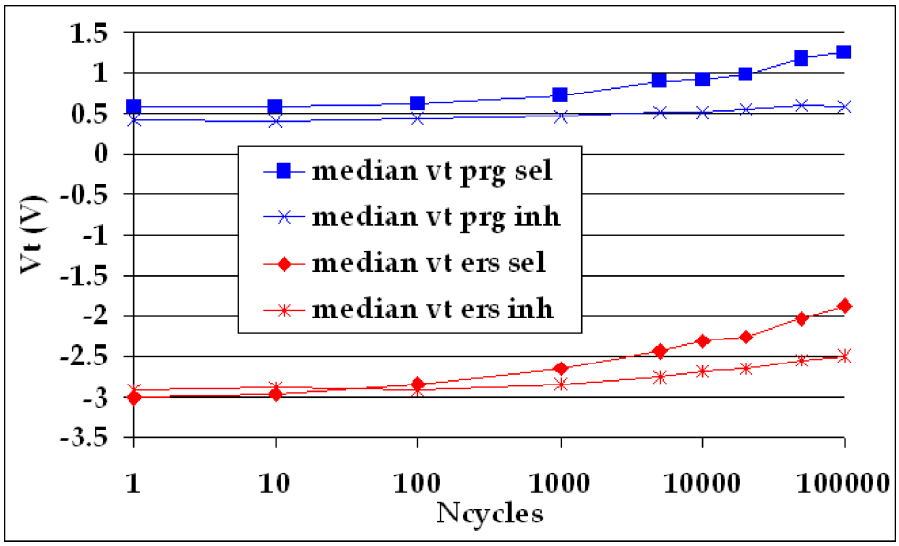


FIG. 4.27 – Un exemple de dégradation des cellules sélectionnées et inhibées au cours du cyclage, avant optimisation du process.

En vue d'identifier le mécanisme responsable de la dégradation des cellules inhibées lors du cyclage de la cellule sélectionnée, uniquement pour certaines conditions process, nous avons mené une étude reposant sur les structures simulées précédemment et sur des mesures sur Silicium. Auparavant, nous allons décrire le phénomène qui permet l'inhibition de la cellule qui reçoit sur sa WL le même signal de programmation que la cellule sélectionnée.

4.7.1 Phénomène d'inhibition en programmation ou "channel boosting"

Afin d'empêcher la programmation de la cellule inhibée lors de la programmation de la cellule sélectionnée, nous utilisons un mécanisme appelé "channel boosting" qui permet de diminuer le champ électrique aux bornes de l'oxyde tunnel et ainsi d'éviter une injection de charges dans la grille flottante [Suh'95][Satoh'99]. Dans notre matrice mémoire en architecture NAND, tous les contacts des cellules sélectionnée et inhibée sont communs à l'exception du contact de Bit Line. Nous rappelons dans la figure 4.28 l'ensemble des polarisations utilisées lors de la programmation de la cellule sélectionnée. Nous voyons que sa Bit Line est polarisée à $0V$ pour amener cette polarisation à l'implant de drain au niveau de la cellule à travers le transistor de sélection BSL, rendu passant grâce à la polarisation de la grille à $4,5V$. Si l'on applique cette même tension de $0V$ sur la Bit Line de la chaîne à inhiber, celle-ci sera programmée en même temps que la cellule sélectionnée.

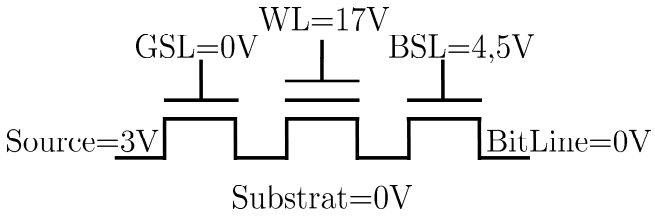


FIG. 4.28 – Conditions de programmation de la cellule sélectionnée

Le phénomène de "channel boosting" est activé en appliquant une polarisation $V_{inh} = 3V$ sur la Bit Line de la chaîne inhibée et en modifiant le signal sur la grille du transistor de sélection BSL qui vaut tout d'abord $4,5V$ puis $1,5V$. La figure 4.29 présente un chronogramme des signaux appliqués sur la chaîne inhibée.

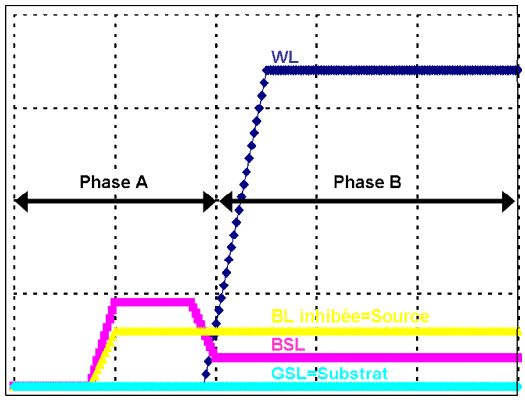


FIG. 4.29 – Chronogramme des tensions d'inhibition

Nous pouvons distinguer deux phases, notées **A** et **B**, que nous allons maintenant détailler.

4.7.1.1 Phase A de précharge

Lors de la phase de précharge, la polarisation $V_{inh} = 3V$ de la Bit Line est acheminée sur le canal de la cellule inhibée via le transistor BSL qui est à l'état passant grâce aux $4,5V$ sur sa grille. Une fois le canal de la cellule préchargé, la tension de BSL passe alors de $4,5V$ à $1,5V$ provoquant une fermeture du transistor de sélection BSL. Le transistor de sélection GSL étant lui aussi bloqué donc la précharge du canal de la cellule mémoire se trouve "emprisonnée". A ce moment-là, nous commençons à faire monter les tensions de programmation de la cellule sélectionnée et la phase **B** débute.

4.7.1.2 Phase B de channel boosting

La WL de la cellule inhibée reçoit également la tension V_{prog}, qui fait monter par couplage capacitif la tension de la grille flottante et entraîne également une hausse de la polarisation du canal et des zones de source et drain qui sont flottants depuis la fin de la phase **A**. La tension de "channel boosting" peut alors continuer d'augmenter jusqu'à $V_{boost} \approx 6,5V$. Le champ électrique aux bornes de l'oxyde tunnel est alors considérablement réduit par rapport au champ dans la cellule sélectionnée pour laquelle $V_{boost} = 0V$.

Au départ de l'étude, nous n'avons pas d'hypothèse privilégiée pour expliquer la dégradation de la cellule inhibée. Les différentes hypothèses de cause de dégradation pouvant être avancées sont les suivantes :

1. une injection de porteurs chauds lors de l'inhibition de la cellule, insuffisante pour programmer de façon visible la cellule qui reste inhibée mais néanmoins suffisante pour dégrader progressivement l'oxyde tunnel et causer une augmentation de la tension de seuil effacée de l'ordre de $500mV$ après 100.000 cycles sur la cellule sélectionnée ;
2. le champ électrique aux bornes de l'oxyde tunnel qui bien que faible permettrait d'injecter quelques électrons dans la grille flottante, suffisamment pour dégrader l'oxyde tunnel et décaler les tensions de seuil avec le nombre de cycles sans pour autant causer une programmation mesurable de la cellule inhibée.

4.7.2 Simulation des conditions d'inhibition

A partir de la simulation complète de notre cellule, que nous avons validée précédemment, nous pouvons désormais réaliser des simulations électriques en vue d'identifier le mécanisme responsable de la dégradation des cellules inhibées. Des études précédentes ont montré qu'il pouvait exister une différence de potentiels V_{DS} entre le drain de la cellule et sa source, causant un effet GIDL[2]. Des électrons sont alors créés du côté de la source de la cellule et sont accélérés par la tension V_{DS} en direction du drain puis sont injectés à travers l'oxyde tunnel [Lee'06].

Les figures 4.30 et 4.31 représentent le potentiel électrostatique simulé dans les zones de source, de canal et de drain de notre cellule mémoire, respectivement avec les mêmes conditions process que nos cellules mesurées et dans des conditions process pour lesquelles apparaît un effet GIDL, très éloignées de nos mesures sur Silicium. Ces potentiels de channel boosting sont obtenus après la montée de l'ensemble des potentiels à leur valeur finale.

[2]Gate Induced Drain Leakage Current

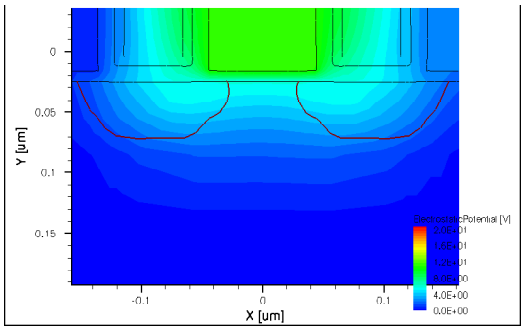


FIG. 4.30 – Simulation électrique de la tension à l'intérieur du canal et dans les implants source et drain de la cellule - sans effet GIDL

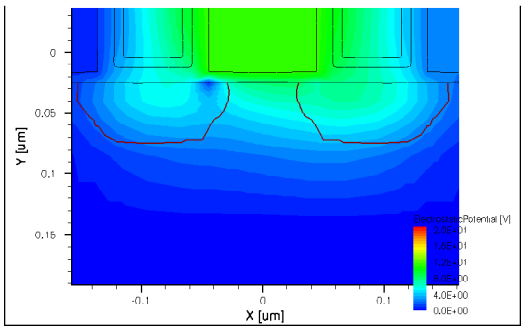


FIG. 4.31 – Simulation électrique de la tension à l'intérieur du canal et dans les implants source et drain de la cellule - avec effet GIDL

Lors de l'ensemble des simulations menées, au cours desquelles nous avons fait varier les doses des implants Source/Drain, leur énergie d'implantation, l'espacement entre la cellule mémoire et les transistors de sélection, nous n'avons jamais constaté de différence de potentiel entre la source et la drain à l'exception de conditions bien particulières d'implant de cellule et d'espacement cellule/transistor de sélection, mais très éloignées de celles correspondant aux structures mesurées sur Silicium. Ce mécanisme ne semble donc pas, si l'on se limite à ces simulations électriques, être celui mis en jeu pour nos cellules.

4.7.3 Mesures de la dégradation en fonction du nombre de pulses élémentaires

Nous avons ensuite cherché à étudier l'impact de la durée des pulses de programmation, pour des polarisations et des temps de programmation totaux identiques, sur la dégradation des cellules inhibées. Nous avons pour cela réalisé des cyclages avec 2 pulses de $100\mu s$, 5 pulses de $40\mu s$ ou 40 pulses de $5\mu s$ en programmation à une polarisation de 17V. La figure 4.32 présente les mesures de la variation de la tension de seuil effacée de la cellule sélectionnée en fonction du nombre de pulses élémentaires de programmation. Nous pouvons remarquer qu'à $50mV$ près, ce qui est de l'ordre de grandeur de l'incertitude de mesure, les variations de la tension de seuil effacée sur la cellule sélectionnée après 100.000 cycles valent $1,2V$ quel que soit le nombre de pulses et ne dépendent donc pas du nombre de pulses élémentaires appliqués. Nous observons en revanche une dépendance logarithmique de la variation de la tension de seuil effacée sur la cellule sélectionnée après 100.000 cycles en fonction du nombre de pulses élémentaires, comme le montre la figure 4.33.

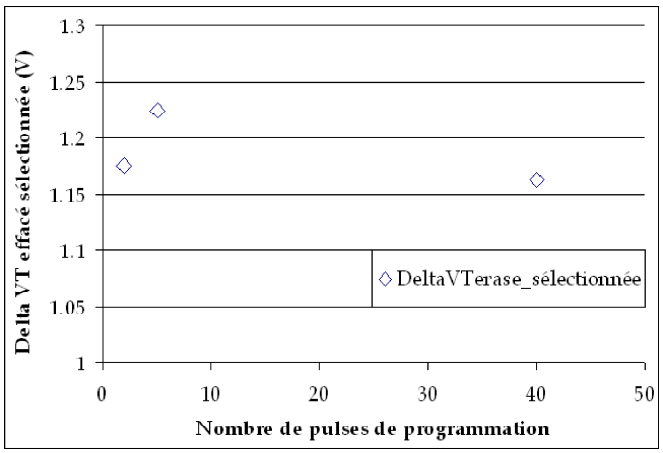


FIG. 4.32 – Variation de la tension de seuil effacée de la cellule sélectionnée avec le nombre de pulses élémentaires, après 100.000 cycles

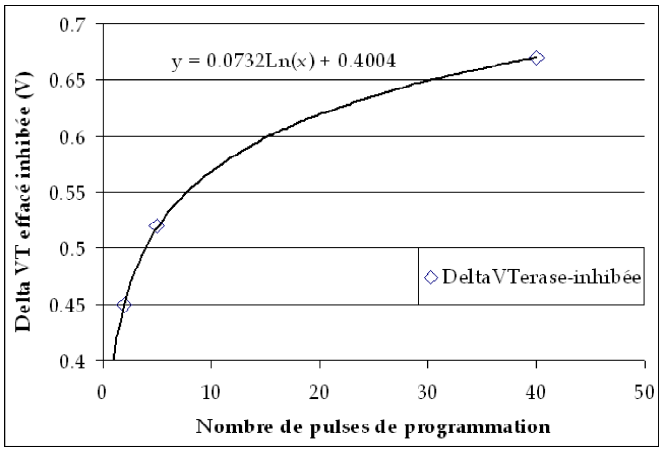


FIG. 4.33 – Variation de la tension de seuil effacée de la cellule inhibée avec le nombre de pulses élémentaires, après 100.000 cycles

Ces résultats de cyclage tendent donc à montrer qu'il existerait un phénomène de dégradation qui apparaît lors des fronts de montée des signaux, mais uniquement sur la cellule inhibée, elle seule étant impactée par la variation du nombre de pulses élémentaires. Nous avons donc cherché à observer ce phénomène en simulation, en se concentrant sur la zone de montée des signaux contrairement à ce que nous avons présenté dans le paragraphe 4.7.2 où les simulations étaient réalisées dans les zones de plateau de la phase d'inhibition.

4.7.4 Simulation des conditions d'inhibition en phase de montée

Nous pouvons donc reproduire, avec les conditions process de nos cellules mesurées, les simulations des conditions d'inhibition lors de la phase de montée des polarisations, afin de tenter de détecter un retard de la montée de la tension de source par rapport à la tension de drain qui causerait une différence de potentiel $V_{ds} > 0$.

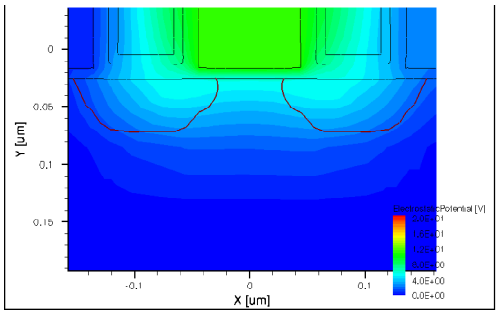


FIG. 4.34 – Simulation électrique de la tension à l'intérieur du canal et dans les implants source et drain de la cellule inhibée à $V_{prog} = 5V$

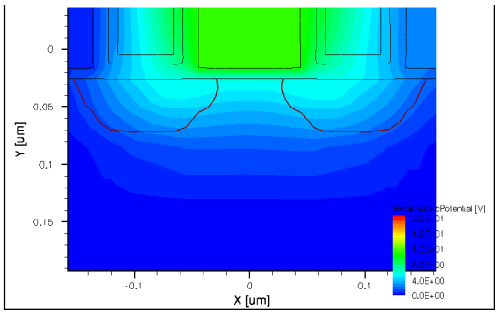


FIG. 4.35 – Simulation électrique de la tension à l'intérieur du canal et dans les implants source et drain de la cellule inhibée à $V_{prog} = 8V$

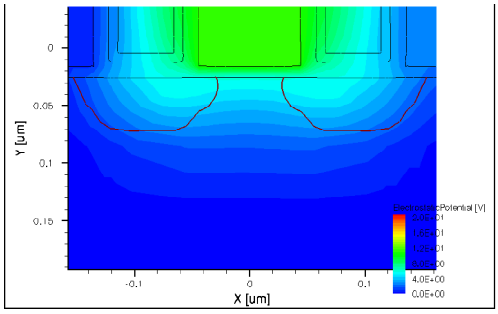


FIG. 4.36 – Simulation électrique de la tension à l'intérieur du canal et dans les implants source et drain de la cellule inhibée à $V_{prog} = 12V$

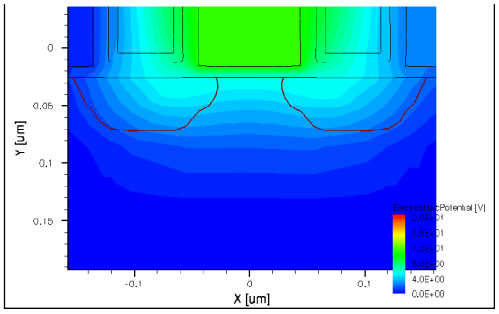


FIG. 4.37 – Simulation électrique de la tension à l'intérieur du canal et dans les implants source et drain de la cellule inhibée à $V_{prog} = 17V$

Ces simulations ne montrent aucune différence de potentiels V_{ds} lors de la phase de montée des conditions d'inhibition, ce qui va à l'encontre des conclusions tirées des mesures du paragraphe 4.7.3. Nous nous sommes donc orientés sur une étude de l'effet du champ électrique sur la dégradation de la cellule inhibée.

4.7.5 Effet du champ électrique sur la dégradation des cellules inhibées

Ne connaissant pas avec précision la valeur de la tension V_{boost} de "channel boosting" dans le canal de la cellule inhibée, nous l'avons déterminée expérimentalement en réalisant des mesures d'efficacité d'injection en programmation, soit avec une tension de 17V sur la grille de contrôle de la cellule inhibée et une tension V_{boost} dans le canal, soit avec une tension V_{boost} forcée à 0V et en cherchant la tension de WL donnant la même efficacité de programmation. Si l'efficacité d'injection est identique dans les deux cas, c'est que nous avons le même courant d'injection et par conséquent le même champ électrique aux bornes de l'oxyde tunnel. La figure 4.38 donne ces différentes courbes d'injection en fonction du temps. En extrapolant sur la figure 4.39 les V_T après $200\mu s$ d'inhibition, nous avons pu identifier la tension à appliquer pour obtenir la même efficacité de programmation comme étant environ égale à $10V$.

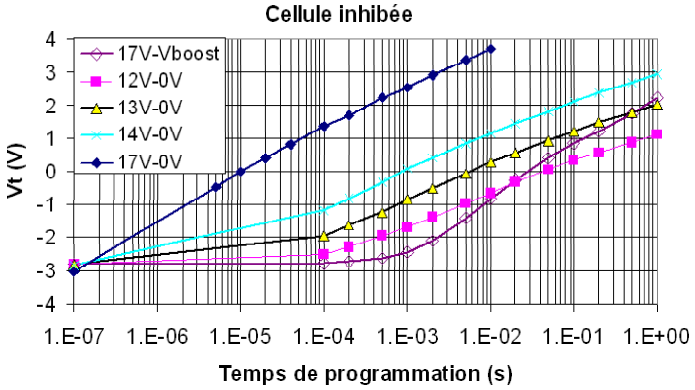


FIG. 4.38 – Détermination de la tension de "channel boosting" par mesure de l'efficacité de programmation

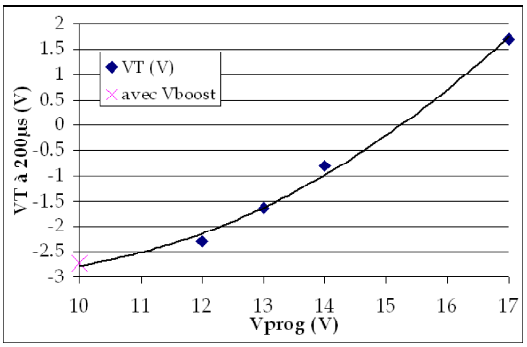


FIG. 4.39 – Extrapolation des V_T à $200\mu s$

A partir de la formule générale de couplage de la tension de grille flottante V_{FG} (Equ. (4.9)), nous pouvons comparer les champs aux bornes de l'oxyde tunnel correspondant aux deux conditions : $V_{prog} = 17V$ et $V_{boost} \approx 6,5V$ ou $V_{prog} = 10V$ et $V_{boost} = 0V$. Nous noterons dans la suite de cette étude les couples de valeurs de V_{prog} et de V_{boost} sous forme mathématique $(V_{prog}; V_{boost})$.

$$V_{FG} = \frac{C_{ONO}.V_{CG} + C_{tun}.V_{boost} + C_{FG_Y}.V_{BSL}}{C_{ONO} + C_{tun} + 2.C_{FG_Y}} \tag{4.9}$$

En prenant $C_{ONO} = 60aF/bit$, $C_{tun} = 40aF/bit$, $C_{FG_Y} = 5aF/bit$ et $V_{BSL} = 1,5V$ pour les deux couples de valeurs $(17V; V_{boost})$ et $(10V; 0V)$, nous obtenons des champs électriques identiques, valant $6,9MV/cm$, pour $V_{boost} = 6V$. Nous pouvons d'ores et déjà faire remarquer que la simulation TCAD bidimensionnelle de la figure 4.30, donne une valeur de "channel boosting" $V_{boost} = 6,1V$, ce qui est en très bon accord avec la mesure qui vient d'être réalisée.

Suite à cette détermination de la tension de programmation donnant avec $V_{boost} = 0V$ le même champ électrique qu'avec une tension de programmation de $17V$ et un $V_{boost} = 6V$, nous avons réalisé des mesures de cyclage sur la cellule inhibée en utilisant ces nouvelles conditions. La figure 4.40 rassemble les décalages obtenus après 100.000 cycles selon les conditions de polarisation sur la cellule inhibée.

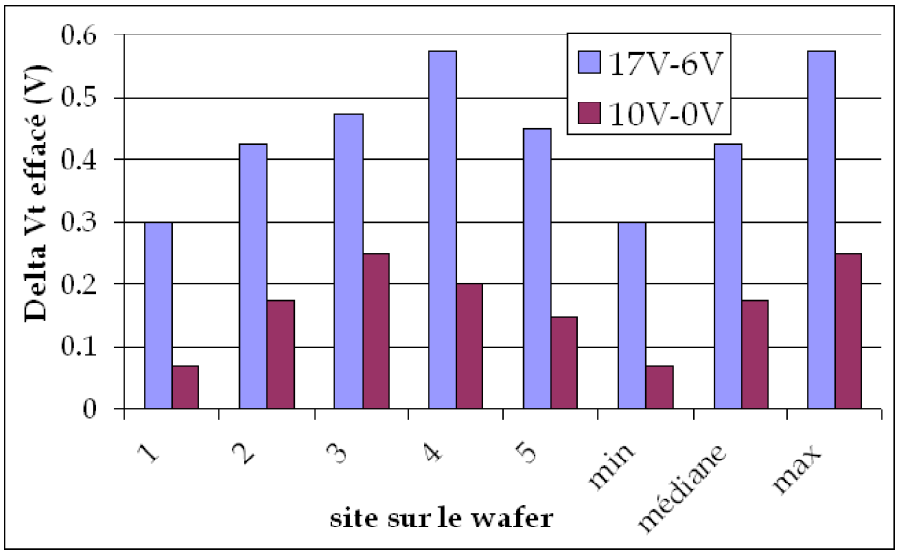


FIG. 4.40 – Comparaison des niveaux de dégradation en cyclage de la cellule inhibée avec $V_{prog} = 10V$ et $V_{boost} = 0V$

A champ électrique équivalent aux bornes de l'oxyde tunnel de la cellule inhibée, le décalage de la tension de seuil effacée pour une tension de programmation de $10V$ est inférieure de $250mV$ à $300mV$ au décalage obtenue avec une tension de

programmation de $17V$. Cela nous permet de démontrer que le mécanisme responsable de la dégradation de la cellule inhibée n'est pas lié au champ électrique et n'est donc pas dû à une injection, même faible, de porteurs par effet tunnel Fowler-Nordheim. Si l'on essaie de déterminer la polarisation de programmation à appliquer sur la cellule inhibée lorsque $V_{boost} = 0V$, pour avoir les mêmes niveaux de dégradations qu'avec une tension de $17V$, nous trouvons expérimentalement une polarisation de l'ordre de $11,5V$, comme le montre la figure 4.41.

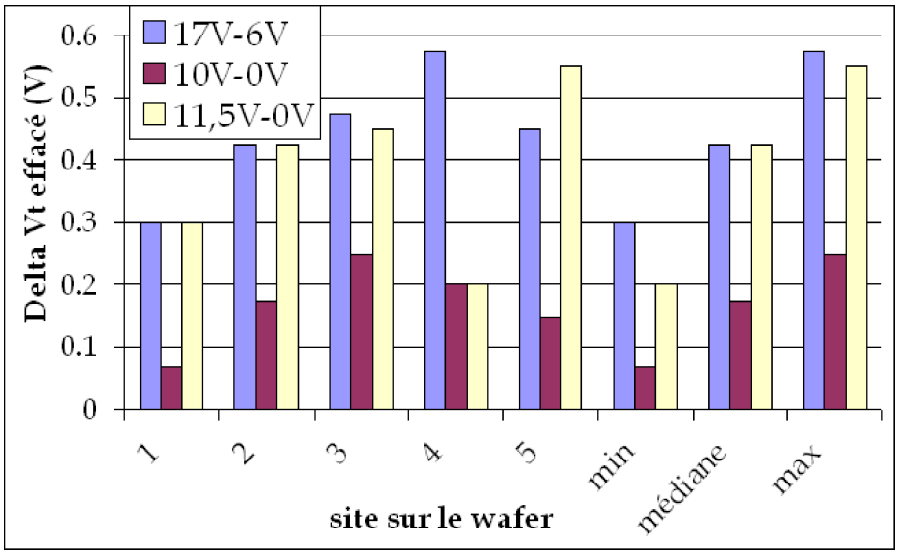


FIG. 4.41 – Comparaison des niveaux de dégradation en cyclage de la cellule inhibée avec $V_{prog} = 11,5V$ et $V_{boost} = 0V$

Ainsi les niveaux de dégradation sur la cellule inhibée sont quasi-identiques entre les deux conditions $(17V; 6V)$ et $(11,5V; 0V)$ alors que les champs électriques, calculés selon l'équaiton 4.9, valent respectivement $6,9MV/cm$ et $7,9MV/cm$. Le champ électrique est dans le second cas plus élevé d'$1MV/cm$, ce qui devrait causer des dégradations nettement plus importantes. La seule cause expliquant les niveaux de dégradation similaires serait qu'un autre mécanisme se superpose à la dégradation par le champ électrique dans le cas d'un channel boosting classique avec $(17V; 6V)$. Cette conclusion nous renvoie vers un phénomène d'injection de porteurs chauds, due à une polarisation V_{DS} non nulle.

4.8 Conclusion

Au cours de ce chapitre sur les perturbations, nous avons évalué les perturbations de grille sur des cellules mémoires Flash en architectures NOR puis NAND. Pour ces deux architectures, nous avons pu mettre en évidence une plus grande sensibilité après

un cyclage à 10.000 cycles. Les valeurs de perturbation augmentent de $100mV$ sur nos cellules en architecture NOR et de $250mV$ dans le cas de nos structures S16 en architecture NAND. Au cours des mesures de cyclage sur la cellule S16, nous nous sommes rendus compte que pour certaines conditions de process, les niveaux de dégradation de la cellule inhibée étaient plus élevés que ce à quoi nous nous attendions avant l'étude. Nous avons donc cherché à expliquer ce phénomène de dégradation. Nous avons pour cela commencé à mettre en place des simulations TCAD bidimensionnelles que nous avons calibrées sur des mesures de tensions de seuil. Pour reproduire le plus fidèlement possible nos conditions réelles, nous avons également développé une simulation tridimensionnelle qui nous a permis d'extraire des valeurs de capacités parasites de couplage entre la cellule adressée et ses cellules voisines. Ces capacités ont par ailleurs été extraites à partir de mesures sur des structures spécifiques, avec une bonne cohérence avec les valeurs simulées. Ces capacités parasites ont ensuite pu être prises en compte dans notre simulation bidimensionnelle, émulant un comportement tridimensionnel. Cette extraction des capacités parasites dans la matrice mémoire comporte un intérêt non seulement pour notre étude des dégradations des cellules inhibées mais beaucoup plus globalement pour l'ensemble des simulations qui doivent être réalisées sur des structures mémoires dans une matrice. A partir de ces structures, nous avons simulé électriquement les conditions d'inhibition par "channel boosting", qui n'ont pas montré d'effet GIDL, ni lors des phases de montée des polarisations, ni lors des plateaux. Cet effet figurait parmi nos hypothèses de départ comme pouvant causer une injection de porteurs chauds à travers l'oxyde tunnel. De plus, nos mesures sur silicium de variation de tensions de seuil lors du cyclage ont à chaque fois semblé mettre en cause un tel effet plutôt qu'une dégradation due au faible champ électrique aux bornes de l'oxyde tunnel. Lorsque nous avons fait varier la durée des signaux élémentaires de programmation, nous en avons conclu qu'un phénomène devait se produire lors de la montée de la phase d'inhibition. Lorsque nous avons déterminé la polarisation à appliquer pour obtenir un champ identique sur la cellule inhibée mais avec $V_{boost} = 0V$, les niveaux de dégradation après cyclage étaient très inférieurs, indiquant qu'un second phénomène se superposait à la dégradation par le faible champ électrique. Nous avons mesuré des niveaux de dégradation similaires avec un champ supérieur de $1MV/cm$ à celui correspondant à une condition d'inhibition classique, obtenue avec $V_{boost} = 6V$. En résumé, nos mesures et nos simulations complètes TCAD ne nous ont pas permis de trancher sur le mécanisme de dégradation responsable de la dégradation des cellules inhibées lors du cyclage de la cellule sélectionnée. Pour cette raison, nous proposerons dans les perspectives de ce manuscrit des mesures complémentaires qui pourront être réalisées.

Références bibliographiques du chapitre 4

[Suh'95] K.D. Suh et al.
"A 3.3V 32Mb NAND flash memory with incremental step pulse programming scheme"
Solid-State Circuits, vol.30, no.11, pp.1149-1156, 1995.

[Satoh'99] S. Satoh et al.
"A novel Gate-Offset NAND Cell (GOC-NAND) technology suitable for high-density and lox-voltage-operation flash"
Proceedings of IEDM, pp.271-274, 1999.

[Lee'02] J.D. Lee, S.H. Hur, J.D. Choi
"Effects of Floating-Gate Interference on NAND Flash Memory Cell Operation"
Elec. Dev. Lett., vol.23, no.5, pp.264, 2002.

[Lee'04] C. Lee, H. Lee, S. Park, C. Park, K. Kim, K. Kim
"Physical Scaling Limit of NOR Flash Memory Cell based on Floating-Gate Interference Effect"
Proceedings of NVSMW, pp.65, 2004.

[Ghetti'05] A. Ghetti, L. Bortesi, L. Vendrame
"3D Simulation study of gate coupling and gate cross-interference in advanced floating gate non-volatile memories"
Solid-State Electronics, vol.49, pp.1805, 2005.

[Lee'06] J.D. Lee, C.K. Lee, M.W. Lee, H.S. Kim, K.C. Park, W.S. Lee
"A new programming disturbance phenomenon in NAND flash memory by source/drain hot-electrons generated by GIDL current"
Proceedings of NVSMW, pp.31-33, 2006.

Conclusion Générale et Perspectives

Le travail réalisé au cours de cette thèse porte sur différents aspects de la fiabilité des Mémoires Non-Volatiles, en particulier des Mémoires Flash. Plusieurs points ont été abordés lors de cette étude :

-La première partie de ce manuscrit traite des méthodes de programmation des cellules mémoires permettant de diminuer la dégradation de l'oxyde tunnel. Une première voie explorée repose sur une étude précédente montrant l'intérêt de l'utilisation de signaux de très courte durée, inférieure au temps de création de pièges stables à l'intérieur de l'oxyde. Sur les technologies que nous avons étudiées, à savoir des structures mémoires en architecture NAND à seize cellules et à une cellule, il ne nous a en revanche pas été possible, probablement en raison de la limitation de la durée minimale des signaux, de montrer un quelconque avantage à l'utilisation de ces signaux. Nous avons ensuite présenté une étude théorique d'optimisation de la forme des signaux de programmation, permettant une injection de charges à courant constant et une réduction du champ électrique aux bornes de l'oxyde tunnel de l'ordre d'$1MV/cm$. Une dernière voie proposée en vue de l'amélioration des méthodes de programmation permettant d'augmenter la tenue des cellules au test en cyclage est le développement et la mise en œuvre d'un algorithme de programmation dit "intelligent" qui garantit une fenêtre de programmation constante avec le nombre de cycles d'écriture/effacement en ajustant le nombre de pulses élémentaires appliqués.

-Dans une deuxième phase, nous abordons les pertes de charges en rétention après cyclage sur des cellules mémoires de type Flash multi-niveaux en architecture NOR. Nous avons pu modéliser l'ensemble de ces pertes par des équations classiques de type Poole-Frenkel et Fowler-Nordheim, chacune représentant les mécanismes prépondérants respectivement en début et en fin de rétention. Nous avons également proposé une modélisation de mesures marginales de "gains" fictifs de charges par un déplacement d'un front de charges se déplaçant du milieu de l'oxyde tunnel vers le substrat.

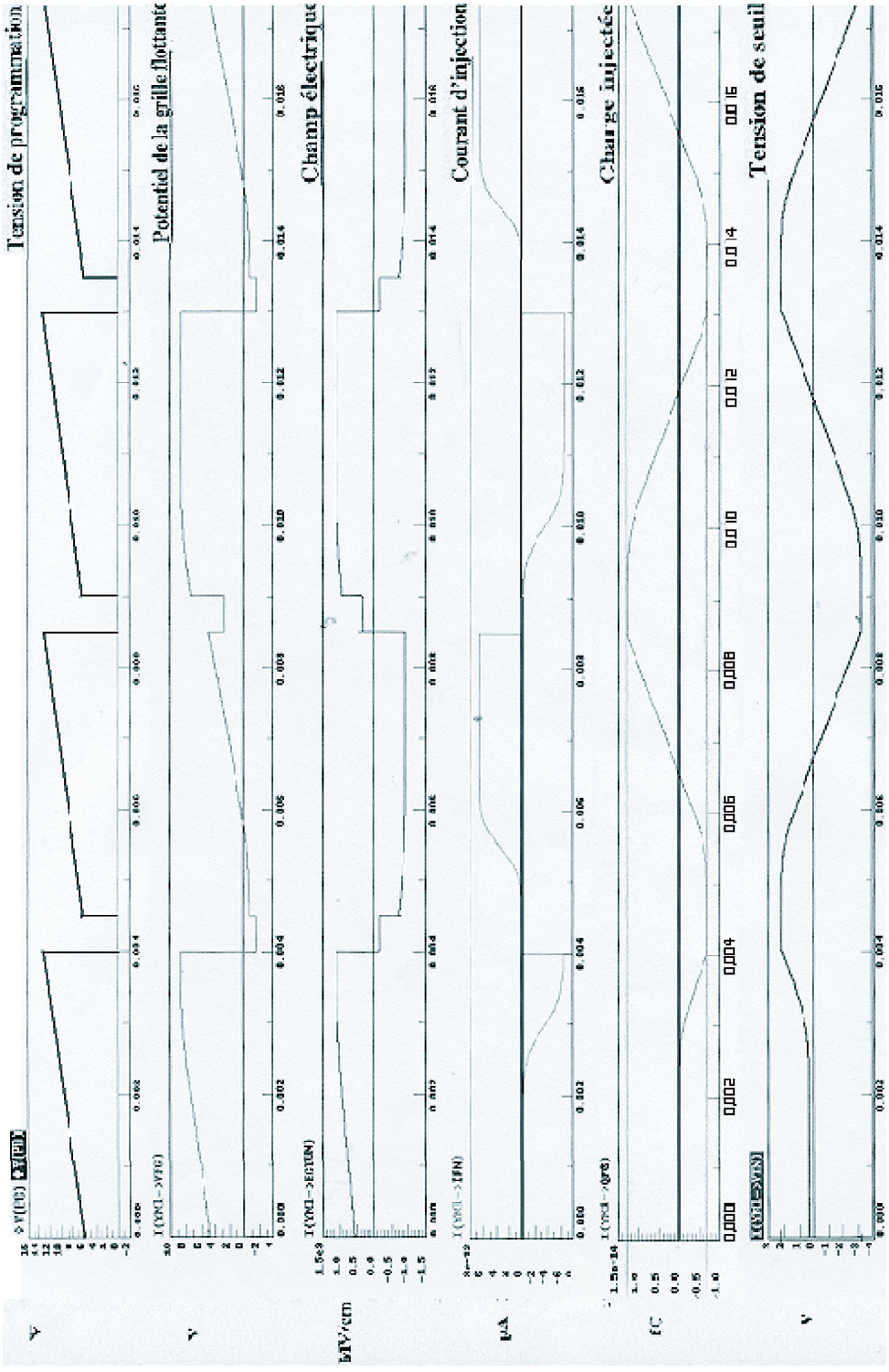
-En dernier lieu, nous étudions différentes causes de perturbations de cellules dans une matrice mémoire, basées sur les polarisations utilisées ou sur les capacités parasites de couplage entre cellules. Nous présentons des mesures de perturbation de grille

sur cellules mémoires Flash en architectures NOR puis NAND à seize cellules, avant et après cyclage pour mettre en évidence l'influence de la dégradation de l'oxyde tunnel sur la tenue à la perturbation. Une dernière partie porte sur l'explication de la dégradation des cellules inhibées lors du cyclage de la cellule sélectionnée. Des simulations TCAD bidimensionnelles ont été menées pour reproduire les conditions d'inhibition utilisées, phénomène appelé "channel boosting". Cette simulation bidimensionnelle a été complétée par une prise en compte des capacités parasites de couplage existant à l'intérieur d'une matrice mémoire NAND. Ces capacités parasites ont été extraites sur des simulations TCAD tridimensionnelles ainsi que par des mesures et des calculs directs. Nous avons trouvé des valeurs cohérentes avec ce qui existe dans la littérature pour des technologies $130nm$. Nous avons ensuite utilisé ces simulations pour tenter de confirmer des mesures de cyclage en fonction de la durée du signal élémentaire qui montre une dépendance logarithmique du niveau de dégradation de la cellule inhibée avec le nombre de pulses appliqués. Les simulations n'ont pas montré de mécanisme d'injection de porteurs chauds lors des phases de montée des polarisations d'inhibition. Nous avons alors mis en place des mesures visant à déterminer expérimentalement la tension de "channel boosting" en recherchant la polarisation V_{prog} à appliquer avec $V_{boost} = 0V$ pour avoir une efficacité de programmation identique, soit un champ électrique identique, à celle obtenue avec $V_{prog} = 17V$ et $V_{boost} \approx 6,5V$. Nous avons ainsi déterminé une valeur de "channel boosting" $V_{boost} = 6V$. Les cyclages réalisés avec ces deux conditions équivalentes en terme de champ électrique ont montré des niveaux de dégradation très inférieurs lorsque $V_{boost} = 0V$. De plus, des niveaux de dégradation comparables ont été obtenus en utilisant un champ électrique $1MV/cm$ supérieur à celui obtenu avec $V_{boost} = 6V$. Ces deux mesures nous ont permis de déduire que la dégradation de l'oxyde tunnel n'était pas directement liée au champ éléctrique appliqué mais qu'un second phénomène, probablement de type GIDL, se superposait et causait la dégradation additionnelle.

Concernant les perspectives de ce travail, nous pourrions renouveler l'ensemble des mesures de cyclage sur des structures avec différentes conditions de process, notamment au niveau des doses et des énergies d'implants Source et Drain, pour mettre en évidence des conditions pour lesquelles la dégradation de la cellule inhibée est maximale. Nous réaliserions alors les simulations électriques sur la structure correspondante pour tenter de mettre en évidence le phénomène incriminé. Cette étude a été commencée en simulation mais la durée des simulations process ainsi que le traitement des multiples réslutats de simulations électriques n'ont pas encore permis d'aboutir à une synthèse. Ce travail pourra être poursuivi lors de la suite de cette année scolaire comme thème de recherche dans le cadre du contrat d'**A**ttaché **T**emporaire d'**E**nseignement et de **R**echerche qui m'a été accordé au sein de l'*IM2NP*. Cette étude pourra ensuite être étendue à d'autres structures mémoires en architecture NAND, actuellement développées par l'entreprise Atmel-Rousset.

ANNEXES

ANNEXE **I** - Simulation Eldo montrant l'injection de charges
à courant constant avec un champ électrique optimisé



Publications

Article dans une revue internationale avec comité de lecture :

1. "Impact of stress on Fowler-Nordheim parameters effects on EEPROM threshold voltage".-**J. Postel-Pellerin**, F. Lalande, P. Canet, S. Boutahar, R. Bouchakour, O. Pizzuto, A. Régnier, Journal of Non-Crystalline Solids,p. 610, 2007.

Colloques et congrès internationaux avec actes à diffusion publique :

1. "Integrated reliability in Non Volatile Memory Cell Design".- Canet P., Lalande F., Razafindramora J, **Postel J.**, Bouquet V., Bouchakour R.- 5th annual Non-Volatile Memory Technology Symposium (NVMTS), Orlando, november 15-17, IEEE Proceedings, p. 66, 2004.
2. "Identification of data retention loss mechanisms on cycled 130nm flash cells".- **J. Postel-Pellerin** , L.Morancho , F. Guyot, T. Pate-Cazal, G. Le Nevez, F. Jeuland, F. Lalande, P. Canet.- 8th Technical and Scientific Meeting of CREMSI FEOL from 130 to 65 nm : scaling challenges, STUniversity, Fuveau, France - October 20-21, 2005.
3. "Impact of stress on Fowler-Nordheim parameters effects on EEPROM threshold voltage".-**J. Postel-Pellerin**, F. Lalande, P. Canet, S. Boutahar, R. Bouchakour, O. Pizzuto, A. Régnier, 6th Symposium SiO2, advanced dielectrics and related devices, Mondello, Italy, June 25-28, 2006.
4. "A full TCAD simulation and 3D parasitic capacitances extraction in 90nm NAND Flash memories".-**J. Postel-Pellerin**, P. Canet, F. Lalande, R. Bouchakour, F. Jeuland, B. Bertello, B. Villard, 9th annual Non-Volatile Memory Technology Symposium (NVMTS), Pacific Grove, USA, November 11-14, 2008.

Conférence nationale avec comité de lecture et acte :

1. "Fiabilité en rétention après cyclage des mémoires non-volatiles de type Flash en architecture NAND".-**J.Postel-Pellerin**, F. Lalande, P. Canet, F. Jeuland, B. Bertello, 10ème Journées Nationales du Réseau Doctoral en Microélectronique, Lille, France, 2007.

Rapports de contrat :

1. "NAND Flash NVM Reliability : Retention after cycling".-**J.Postel-Pellerin**, F. Lalande, P. Canet, F. Jeuland, B. Bertello, Rapport annuel du projet EREVNA ATMEL-L2MP, Rousset, France, 2006.
2. "NAND Flash NVM Reliability : Degradation of inhibited cells".-**J.Postel-Pellerin**, F. Lalande, P. Canet, F. Jeuland, B. Bertello, Rapport annuel du projet EREVNA ATMEL-L2MP, Rousset, France, 2007.

Résumé

Cette thèse étudie divers aspects de la fiabilité des mémoires, notamment les tests en endurance et les tenues en rétention sur des mémoires Flash, en architectures NOR et NAND. Nous abordons différentes méthodes de programmation existantes dans la littérature, à savoir l'utilisation de signaux très courts et un algorithme de programmation intelligent, que nous avons appliquées sur nos cellules mémoires afin de réduire la dégradation qu'elles subissent lors des phases successives de programmation/effacement. Les améliorations observées n'étant pas significatives, nous n'avons pas choisi d'utiliser de tels signaux dans la suite de notre étude. Nous présentons également une théorie des signaux optimisés qui n'a pas été approfondie ici mais que nous avons étudiée dans une étude préalable à cette thèse. Nous présentons ensuite une modélisation des pertes de charges en rétention à partir d'équations simples de types Fowler-Nordheim et Poole-Frenkel qui se superposent et respectivement prépondérantes à des temps de rétention élevés ($t > 200h$) et courts ($t < 200h$). Nous proposons enfin une études des perturbations intervenant dans une matrice mémoire, à la fois du point de vue des tensions électriques appliquées sur les cellules mais aussi du point de vue des capacités de couplages parasites. Nous avons dans un premier temps évalué les valeurs de perturbation de grille sur des cellules mémoires Flash en architecture NOR puis NAND avant de traiter des capacités parasites entre cellules dans une matrice. Nous avons été amenés à étudier ces capacités dans la cadre de l'étude des dégradations excessives des cellules inhibées lors de tests en endurance pour certaines conditions process non-optimisées. Nous avons pour cela développé une simulation TCAD bidimensionnelle à partir des étapes process réelles que nous avons ensuite calibrée sur des mesures sur silicium. Enfin cette simulation a été complétée par une prise en compte des capacités parasites de couplage, extraites sur une simulation tridimensionnelle d'une matrice 3x3 de cellules mémoires. Les valeurs de ces capacités ont été validées par des mesures sur des structures de test spécifiques et par calcul géométrique. Notre simulation bidimensionnelle émule donc un comportement tridimensionnel tout en restant dans une rapidité de calcul liée à une simulation 2D. Nous avons ainsi pu développer des simulations électriques permettant de visualiser le phénomène d'inhibition des cellules, tout au long de l'application des diverses polarisations sur la structure.